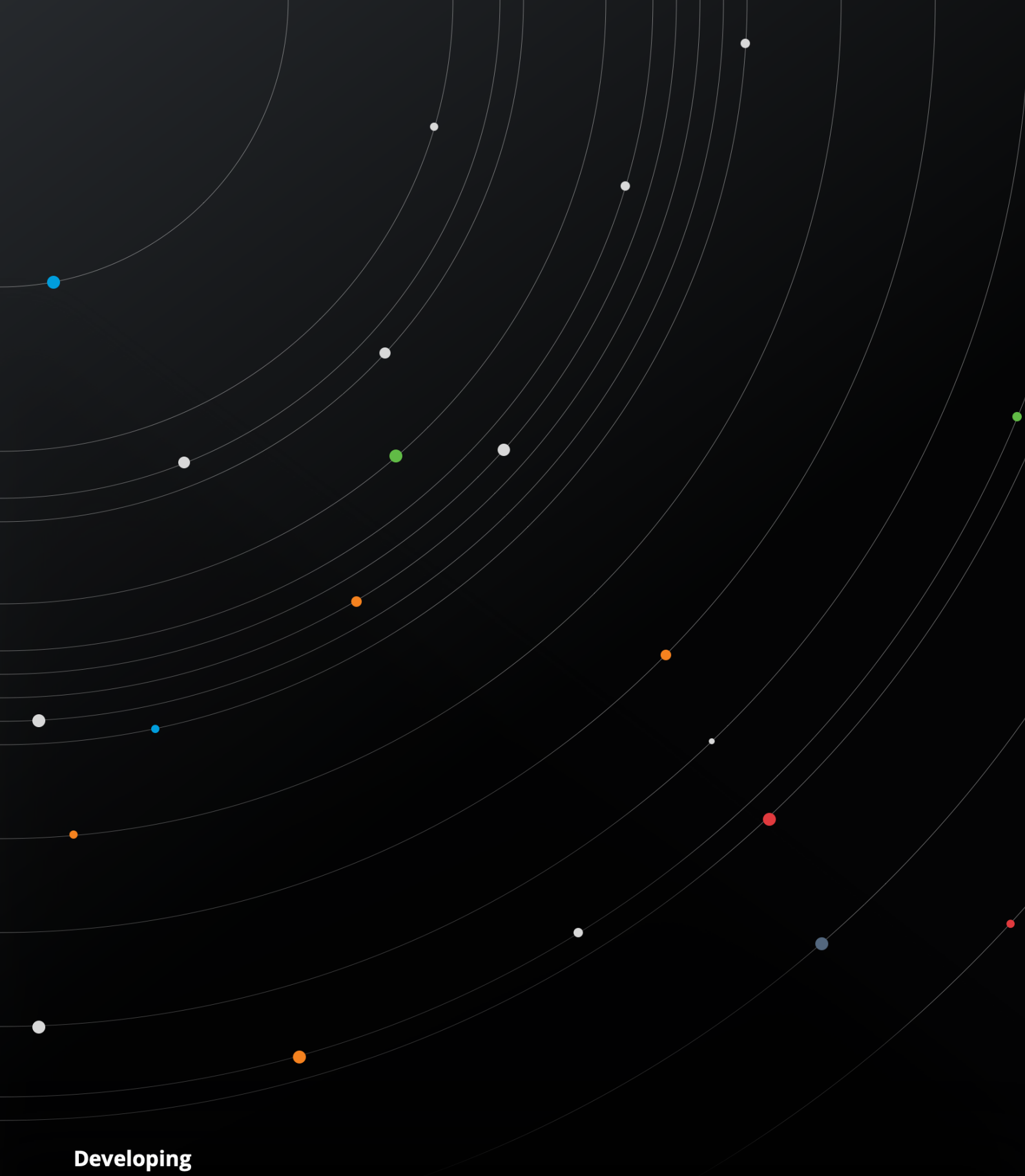


Developing data-driven tools

to minimize data complexity in transport geography research

Filipe Teixeira

"The time needed to acquire data from any astronomical object increases at least as quickly as the square of the distance to that object, so any service that can accumulate custom ensembles of already captured images and data effectively brings the night sky closer." - Jim Gray

Developing data-driven tools to minimize data complexity in transport geography research

Filipe Alberto Marques Teixeira

Proefschrift aangeboden tot het behalen van de graad van Doctor in de Wetenschappen: Geografie (UGent) (2020)

"We call it science, but in the end, it's just unbridled curiosity."
Neil deGrasse Tyson

Supervisors

Prof. dr. Ben Derudder

Department of Geography, Ghent University

Prof. dr. Mario Pickavet

Internet Technology and Data Science Lab, Ghent University

Members of the Examination Committee

Prof. dr. Veerle Van Eetvelde (Chair)

Department of Geography, Ghent University

Prof. dr. Ate Poorthuis

Department of Earth and Environmental Sciences, KU Leuven

Prof. dr. Frank Witlox

Department of Geography, Ghent University

Prof. dr. Kurt Fuellhart

Department Geography & Earth Science, Shippensburg University

Em. prof. dr. Philippe De Maeyer

Department of Geography, Ghent University

CONTENTS

Acknowledgments

Whereas a PhD takes in theory about four years to complete, the path leading to it and the people involved in the process can span for much longer than that. Every researcher perceives and experiences the process leading to achieving PhD and receiving the traditional 'funny hat' differently. However, beyond the academic and technical challenges I have faced as a newcomer to the world of Geography, I have dealt with loss and struggled with my own mental health. Thankfully, I have been supported by friends and colleagues without whom I would never have been able to achieve what I have achieved so far. In Portugal, we refer to our PhD supervisor as "orientador" or "mentor", both referring to someone who guides you or mentors you throughout an academic process. In Dutch they refer to the word "promotor", while in German the word "doktorvater" is curiously used as it translates to the PhD's father. When I first sat with Ben to talk about a potential collaboration, I was sceptical about going back to academia, and had almost no hope that someone would believe in my reintegration into the scientific world after 6 years away from it. However, Ben believed that a Portuguese Biochemist who could not code in R and knew little about spatial analysis, would be able in four years to develop the necessary skills to get where we are at the moment. It was not always a smooth path and we did not always agree with each other. Nevertheless, I will be forever thankful for believing in me by giving me this opportunity. But most importantly for pushing my boundaries further than I could imagine, by challenging me throughout this process.

Within the realm of supervisors, I would like to specially thank to Kurt Fuellhart. Our long skype brainstorms nurtured the scientist in me and made me want to break barriers I would otherwise not have thought of. Your passion for the subject and your drive for science, energised me whenever I thought I was bound to fail. Finally, a special thanks to Frank Witlox and to the Social and Economic Geography (SEG) group, for giving me the space, structure and opportunity to pursue some of my endeavours in research. A big and special thank you to Mario Pickavet and Pieter Audenaert, for their help in the beginning of the project and for Mario's valuable feedback prior to finalizing this dissertation.

A PhD however, brings more than just academic supervision or stimuli. With that in mind, I first have to thank Freke Caset. Life is a strenuous path similar to the mountains you showed during the defence of your own dissertation. However, the same could be easily said about our friendship. We did not always agree with each other, we did not always communicate perfectly, but I would not be here without you. I will be forever thankful for all the times you broke through my stubbornness, and for all the other times you were here for me in some of the darkest moments in my life. More than just a colleague, you are a true friend. To Virginia, "muchas gracias por todo". You surely more than just brightened these past few years with our nice conversations and with your support. It is always hard to explain how the warmth from our countries is so much needed in life. I am more than just glad and thankful for you being here for me. Our office would not be the same however, without Jorn. Thank you for all the laughs and for your extraordinary Frans Bauer impressions. Galuh, you have taught me more than what a fellow scientist can teach. As a colleague and as a friend you shared kindness and wisdom throughout our conversations. I am extremely happy that our paths crossed in Belgium and I hope to see you again wherever we are. When talking about kindness I cannot forget to mention Melakuh. I will always cherish our brief but insightful talks. The SEG is not just the people I have mentioned. A big thank you for all those who throughout these years had an impact in my life one way or another. A special shout out and an immense thank you to Bart de Wit and Luc Zwartjes for their assistance and patience in helping me

with my data. As someone who started with no knowledge at all in GIS, things would have been much harder without your help.

But a PhD is not simply made of colleagues and academic staff. In Belgium the team supporting the academic staff (i.e. 'administratief en technisch personeel') are constantly running everything on the background, assuring that things run smoothly. With that in mind, I would like to first thank Sofie De Winter. Your commitment, kindness, empathy and dedication are an absolute true inspiration for anyone. You went far and beyond to make sure that all of us had things running smoothly, being when we were in Italy for a conference or just in need of help at work. No words are enough to thank you for everything you have done for me and for our colleagues every day at work. Paul Schapelynck, I cannot thank you enough either for every time you helped me with any request I had. Your friendliness and joyfulness every day were truly contagious. Wim Van Roy and Steven De Vriese, your technical support is absolutely invaluable for our department. During the four years I was there for, you were always available to help even when you were extremely busy, or even when something required extra efforts. A special thanks to Helga for your amazing skill and determination to solve the many bureaucratical challenges Belgium is famous for. A big thank you as well to Karine Van Acker, for all the help and interesting conversations we had about photography. There are many other names to be mentioned but that does not mean you are any less valuable than the ones I wrote above.

Some might disagree but there is a life outside of academia. First of all, I would like to thank my parents for "teaching me" science. From the lessons of Carl Sagan to Jacques Cousteau, to all the books from "Once upon a time... Life", you have fostered the scientist in me. To my mom for giving me the platform to being able to study, no matter how absurd my curriculum changes were. To my dad for introducing me to the arts and literature and for fostering my relationship with nature and with the mountains I much love right now.
Liza Notaerts I would never have started this PhD Without you. Your support, love, kindness and dedication made me the person I am today and gave me the strength and confidence to embrace this challenge. You were my rock for almost 5 years, and I will never forget that. A big thank you as well to Peter and Sas. All your support in darker times, together with all the good moments you both gave me, are something I will never forget. "Merci et à bientôt". Ari and Adelaide. Thank you for bringing the most needed South European warmth. Thank you for all the amazing food and drinks we have shared for the past few years and thank you for all your support and help.
A big thank you to Ania. For the past year and a half, you have played an extremely important role in my life which cannot simply be ignored. Thank you for your support during these hard times, the times you made me laugh and the many times you critically judged my food creations. I could have not asked for a better friend. Seinfeld will never be the same without you. A big thank you as well to Olle Abrahamsson and to Isaac Levi Henderson. Part of going to conferences and courses is the people we meet there. You guys made those not always so exciting conferences and courses, definitely brighter and interesting. To Diogo a big thank you for your friendship and all the times we exchanged series, movies and games, despite the geographic distance. Ricardo Maia, a big thank you for all the nice chats and adventures even when so far from Belgium. I'll always look forward to chatting about photography and tech. Ricardo Nabo, a big big thank you. All the amazing adventures we had together and all the times you made me laugh, surely compensate the many other times you almost killed me with your driving. Lisa Van Coillie thank you for showing me kindness and for showing me that it's in the simple things that we can help others. Being some heart-warming food, or a walk at night in Bruges.

Pieter-Jan de Schryver, a thank you is surely not enough to express my appreciation for your work. Your support throughout these past 6 years goes beyond any words I can write in this very short section.

Before I finalize this section, I would like to briefly talk about mental health. While it is impossible to cover such complex topic in only a few lines of text, I will quote one of my favourite comic book artists (i.e. Matthew Inman). "Maybe I'm just built differently. Maybe I was born anxious and angry and this is how I find peace with the universe. Maybe I truly am miserable and everyone else is feeling something I'm not. Or maybe they're all full of shit. It's irrelevant. Because I'm not happy and I don't pretend to be. Instead, I'm busy. I'm interested. I'm *fascinated*. I do things that are meaningful to me, even if they don't make me 'happy'. I run. I run fifty miles at a time. I run over mountains until my toenails fall off. I run until my feet bleed and my skin burns and my bones scream. I read. I read long, complicated books about very smart things. And I read short, silly books about very stupid things. I read until their stories are more fascinating to me than the people actually around me. I work. I work for twelve hours a day. I work until I can't think straight and I forget to feed myself and the light outside dims to a tired glow. I work until I smell weird. When I do these things, I'm not smiling or beaming with joy. I'm not *happy*. In truth, when I do these things, I'm often suffering. But I do them because I find them meaningful. I find them compelling. I do these things because I want to be tormented and challenged and *interested*. I want to build things, and then break them. I want to be busy and beautiful and brimming with ten-thousand moving parts. I want to hurt, so that I can heal. I'm not unhappy. I'm just busy. I'm *interested*. And that's ok."
Discussions about mental health and happiness tend to be either avoided or simply buried in the technical and medical aspects of it. In fact however, mental health is probably one of the most difficult and yet simple topics to discuss. This dichotomy between being a complex and often unmeasurable topic, and the simplicity of raising awareness to it, should not in any sense stop us from discussing it. My point here is that if you see someone struggling with their mental health, just stop and listen. Most things are easy to handle. Most problems are easy to approach even when professional help is needed. In the end we are all as Matthew Inman said, "perfectly unhappy", and that's pretty much ok. After working in different fields of science, I still struggle with what comes from it, but that is as well what makes me thrive and love what I do. In any case there is only one thing I can recommend for those who love science and yet struggle with it.

Be busy.
Be interested.

Finally, this is dedicated you Sophie. I would send you a copy, but post does not send packages to wherever you are. One of your favourite quotes from Bill Watterson, is still what drives me to explore in science: *"It's a magical world, Hobbes, ol' buddy... Let's go exploring!"*.

List of Figures

List of Tables

List of Abbreviations

AI – Artificial Intelligence
API – Application Programming Interface
BTS – Bureau of Transport Statistics
DB1B – Origin Destination Survey
fMRI – functional Magnetic Resonance Imaging
GIS – Geographic Information System
GUI – Graphical User Interface
GWA – Genome Wide-Association Studies
IATA – International Air Transport Association
MAR – Multi-Airport Region
OAG – Official Airline Guide
OD – Origin Destination
OS – Operating System
OSM – Open Street Maps
PSS – Planning Support System
T-100 – Air Carrier Statistics – form 41 traffic
TOD – Transit Oriented Development

CHAPTER

1

Introduction

"The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths." - Ada Lovelace

1.1 Outline and objectives

"The time needed to acquire data from any astronomical object increases at least as quickly as the square of the distance to that object, so any service that can accumulate custom ensembles of already captured images and data effectively brings the night sky closer. (...) Using Microsoft's WorldWide Telescope, anyone can pan and zoom around the sky, at wavelengths from X-ray through radio, and anyone can navigate through a three-dimensional model of the Universe constructed from real observations, just to see what's there. Anyone can notice an unusual correspondence between features at multiple wavelengths at some position in the sky and click right through to all the published journal articles that discuss that position."—Jim Gray (The Fourth Paradigm)

This dissertation deals with the need for, and the key role and development of, data-driven tools and methodologies for minimizing data complexity in transport geography research. It does so by developing a diverse range of analytical tools that show that this is both feasible and useful. In addition to developing these tools in the strict sense, in the dissertation I also examine the various challenges in the process from data acquisition to curation, analysis, and visualization. In other words, my overall objective is to develop and reflect on data-driven tools and methodologies that can minimize data complexity in transport geography research. Although this dissertation focuses on and contributes to topics typically studied in transport geography (e.g., air transport, transport planning support systems), the main topic of this thesis (i.e., data-driven tools, data democratization) has more general implications and therefore has ramifications beyond transport geography. For the time being, I will straightforwardly use the concept 'data-driven tools' to refer to software developed to collect, process, analyze, or visualize data. However, I will refine this straightforward definition so as to more precisely specify its remit.

1.1.1. The fourth paradigm

Over the past few years, there has been a steady rise in the availability and volume of data in general and spatial data in particular (Miller & Goodchild, 2015). The abundance of data containing either explicit or implicit spatial information may be leading to what some are calling the fourth paradigm of science (Hey et al., 2009). Before this fourth paradigm was posited, science was founded on three paradigms: experimental, theoretical (with Kepler's Laws or Newton's Laws of motion), and simulation (when the theoretical models grew too complex to be solved analytically). However, the volume of data generated soon became impossible to be analyzed by classical simulation models, thus giving birth to this fourth paradigm. Gray (2005) claims that this new paradigm is driven by a flood of observational data that risks overwhelming scientists. Considering these fundamental changes in the nature and volume of available data, the processes involved in data acquisition, curation, analysis, and visualization continue to consist of complex paths in many research domains (Gray et al., 2005). Today, the challenges associated with scientific data analysis involve a considerable demand for tools and computation resources to put scientists in control of their data (Gray et al., 2005; Hey et al., 2009). Some of these challenges, which may include large datasets; complex data in terms of size, heterogeneity, or dimensionality[1]; inadequate or complex off-the-shelf tools that require years of training; etc. collectively lead researchers to having to make compromises throughout their research projects. For example, a large dataset that spans decades might force researchers to narrow their

[1] High dimensional data is often characterized by a high number of variables, making data difficult to analyze using conventional methods (e.g. multivariate analysis).

focus to one year's worth of data due to the computational demands behind parsing and processing terabytes of information.

1.1.2. Data democratization

Alongside the fourth paradigm of science and the flood of observational data, there has been a movement calling for a democratization of data. This concept refers to unlocking data that was previously only accessible by scientists with the necessary skills to access and manipulate that data. The democratization of data then refers to a "non-technically skilled" user of information systems being able to, in principle, access and analyze data. In principle, this would make all kinds of data resources available to everyone regardless of their IT skills. In this dissertation, I focus on data democratization within a scientific context. I argue that, by making data broadly accessible to use, analyze and visualize, we can empower researchers without a technical (e.g., programming) background with data that would otherwise be difficult to use. However, the democratization of data does not necessarily exclusively focus on researchers lacking a technical background. Some well-known examples include Google Maps,[2] Waze,[3] and UBER,[4] where citizens are given access to a vast range of data otherwise inaccessible to the non-technically skilled. Google Maps, for example, which borrows concepts from Geographic Information Systems (GIS), has for the past few years become part of the daily lives of many.

In theory, data democratization allows more people to gain access to data, thus allowing them to be able to make more data-driven decisions. However, there are some papers (Gray et al., 2005; Hey et al., 2009; Fahey, 2014) critically reflecting on the drawbacks of widespread access to data. One of the most common arguments is that when data is accessed by people without a strong and suitable background (i.e., data science, statistics, IT), data are more likely to be misinterpreted (Fahey, 2014). With this in mind, in this dissertation I call for data democratization in transport geography, with a special focus on air transport data and Multi-Airport Regions (MARs) on the one hand, and on transit-oriented development (TOD) research on the other hand.

In the case of MARs, I focus on making the two most used data sources (e.g., air transport data, traffic data), more accessible by, first, proposing a format that allows immediate scrutiny, and second, developing a methodology that allows exploring the spatio-temporal dynamics within MARs. For the past decades, airports have become key infrastructure in connecting cities and regions in an increasingly complex and integrated global economy, making the integrated planning and management of their accessibility of key importance. Defining a region as having multiple airports within the context of MAR represents a major shift from the past, when it could be assumed that airports functioned within more or less clearly defined catchment areas. In a MAR, airports are part of a more complex system where there are different airports with overlapping and interacting catchment areas. The significant impact of MARs on economic development and on cities and their regions (O'Connor & Fuellhart, 2016) has made it essential to understand some of the dynamics these areas generate as well as their potential for regional and global connections at the global scale (Fuellhart & O'Connor, 2019). With this in mind, MARs become an attractive research topic.

[2] https://maps.google.com
[3] https://www.waze.com
[4] https://www.uber.com/be/en/

There is a solid body of MARs research in the United States, largely using datasets provided by the Bureau of Transport Statistics (e.g., Fuellhart et al., 2013, 2016; Neal, 2014). The MAR concept is thereby becoming more broadly established, especially in the United States geographical context: regional codes are used in booking systems and sometimes explicitly recognized by the International Air Transport Association (IATA), as shown, for example, by the single codes used for the San Francisco Bay Area (QSF) and New York (NYC) airports. Even in regions that may not be on the map of 'classical' MARs, MAR-like research questions emerge. Fuellhart (2007), for example, showed airport substitution patterns for Harrisburg International Airport (MDT), located in south-central Pennsylvania, towards various proximate airports. The consistency in MARs research allows the development of a narrative that explores the challenges in the processes from data gathering to curation. This can be achieved by focusing on the Bureau of Transport Statistics (BTS) Origin Destination Survey (DB1B) and Air Carrier Statistics Form 41 Traffic (T-100) datasets. This is supported by the fact that MARs research also often uses a broad and very diverse range of complementary data sources and types (e.g., road traffic, air fares, connectivity) to measure airport accessibility and utility and subsequently model airport choice in these areas (Pels et al., 2003; Ishii et al., 2009; Mun & Teraji, 2012; Fu & Kim, 2016).

The challenge of making data accessible does not end with data parsing and analysis. In StationsRadar, I tap into the field of visual analytics and accessibility instruments to develop a data-driven web-based tool. Through a series of workshops directed to planning and policy stakeholders, together with my colleague Freke Caset, I have developed an interactive web data-driven tool to support integrated land use and transport strategy development at railway station locations. StationsRadar differs from other accessibility instruments, as it was developed in close dialog with policy and planning stakeholders after a series of workshops. This approach will be fully described in Chapter 5.

In the remainder of this introductory chapter, I first discuss the rise in the availability of data in transport geography, and the subsequent challenges in acquiring, curating, analyzing, and visualizing data (Section 1.2). This is subsequently taken as the starting point for elaborating on the concept of "making big data small" (Poorthuis & Zook, 2017) and introducing some concrete examples of data-driven tools and methodologies used to democratize data in the context of transport geography (Section 1.3). Against this backdrop, the various formative chapters in this dissertation examine four of the stages from data acquisition to visualization: (1) acquiring and parsing data with SKYNET, an R package[5] that allows generating bespoke air transport statistics from a freely available dataset; (2) curation of the dataset (i.e., BTS DB1B) by using SKYNET to identify biases and therefore validate the data to be used in subsequent research; (3) after capture and curation, I analyze the data to acquire insights on the spatio-temporal dynamics of the New York MAR; and finally, (4) I combine the knowledge and tools developed in stages 1 to 3 in a somewhat different context: to create an open and interactive planning support tool, intended to support integrated strategy making for railway stations and their surroundings in the regions of Flanders and Brussels, Belgium. Before moving to these discussions, I briefly summarize the major tenets of each chapter in their own right.

[5] Programming language and free software environment (RStudio) for statistical computing (https://www.r-project.org/about.html).

Chapter 2: SKYNET—An R package for generating air passenger networks for urban studies

Filipe Teixeira	Conceptualization, formal analysis, investigation, methodology, software development, visualization, writing
Ben Derudder	Article review, editing

There is a long tradition of urban studies invoking air transport data either for tackling the city/air transport-nexus directly (e.g., in research on the causality between urban-economic development and air transport connectivity) or as a secondary data source (e.g., in research mapping city networks).

However, air transport statistics rarely come in a format that allows for their immediate scrutiny in light of the research questions at hand, so handling and transforming these data often involves both practical challenges and considerable effort. With resources often being limited (e.g., time constraints, IT infrastructure), air transport researchers are confronted with having to choose between expensive but structured datasets (e.g., Official Airline Guide[6]) and freely but harder to parse datasets (e.g., BTS DB1B). Commercial datasets often simplify the process from acquiring data to having it in a format that allows researchers to draw conclusions by means of a simplified web-based tool or Application Programming Interface[7] (API). Meanwhile, freely available datasets often present data as a collection of CSV[8] files, leaving the curation, analysis, and visualization to the researcher. However, despite the obvious demand for tools that allow for the manipulation of freely available datasets, the offer remains circumscribed to a few options that are limited in functionality. Some of those tools include: R "cansensus," which provides access to Statistics Canada's Census data, R "acs," which downloads, manipulates, and presents the American Community Survey data, and "AIRNET," a program for generating intercity networks from the BTS DB1B and T-100 dataset.

Against this backdrop, I introduce "SKYNET," a flexible R package that allows generating bespoke air transport statistics for urban studies based on publicly available data from the BTS in the United States. The basic elements of the package are explained, after which I demonstrate its usefulness by showing its potential for addressing research questions emerging in the literature on 1) evolving urban landscapes of air travel accessibility, and 2) differences in intercity air transport networks by scale, types, and season. I argue that this R package has the potential to become the backbone of a range of easily navigable tools overcoming some of the main methodological challenges researchers face when handling complex airline data in an urban context.

Chapter 3: Revealing route bias in air transport data: The case of the Bureau of Transport Statistics' (BTS) Origin-Destination Survey (DB1B)

Filipe Teixeira	Conceptualization, formal analysis, investigation, methodology, software development, visualization, writing
Ben Derudder	Article review, editing

[6] Flight database and statistics (https://www.oag.com).
[7] Functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service.
[8] Comma separated values

The purpose of this chapter is to explore how potential biases in air transport datasets can be revealed and detailed. Here, the DB1B dataset, a freely available dataset provided by the US BTS, is used as the focus of the study. For the past 20 years, air travel has been dramatically increasing, with worldwide passenger numbers more than tripling from 1.025 billion in 1997 to 3.227 billion passengers in 2017, with the IATA forecasting these numbers to double again by 2036. It is important, however, to stress the impact of the current COVID-19 situation on air transport. For example, according to the BTS, the number of monthly passengers changed from approximately 61 million in January 2020 to 2 million in April 2020, following the travel restrictions imposed by the US government. The long-term effects of this pandemic on air travel are still uncertain. The evolution of the current situation is still hard to forecast, mostly due to the unpredictable nature of the virus and the novelty of the situation. Nonetheless, the growing impact of air transport has clearly fueled recent waves of air transport research (Ishutkina & Hansman, 2008; Air Transport Action Group, 2010). However, despite the growth and diversification in data (Poorthuis & Zook, 2017) challenges remain, partially because of the uneven availability and formatting of data (Derudder & Witlox, 2005b). Another challenge air transport researchers face is in the limited choice of data sources to use. The core challenge however, remains in the data collection where there are two approaches available. The first one involves building a database from scratch, by acquiring web-based travel data to access a meta-search engine or an online route planner (Polidoro et al., 2015; Lieshout et al., 2016). The second approach involves a primary dataset and the tools already included with them. This can vary from freely available data (e.g., DB1B, T-100), to often expensive but well-established data (e.g., OAG). The DB1B dataset comprises a 10% sample of reported tickets for US domestic flights. The T-100 dataset, on the other hand, which I use to validate the DB1B, represents a full dataset containing domestic and US-related international airline market and segment data. The most immediately noticeable difference between the DB1B and the T-100 lies in what they represent. Whereas the DB1B shows reported tickets and is grouped per quarter, the T-100 reports flights grouped per month. Despite their US focus, both datasets are widely used in the air transport literature (Fuellhart, 2007; Neal, 2014b; Roucolle et al., 2020). However, to the best of my knowledge, the methods for data collection, sampling, and overall quality of the DB1B dataset have rarely been scrutinized (Boyd & Crawford, 2012; Poorthuis & Zook, 2017).

To explore potential bias in air transport datasets, this chapter follows the four stages of analysis and assessment of the BTS datasets (i.e., DB1B, T-100): describing the data, testing the randomness of the 10% sample, exploring the overlap between the two datasets, and assessing the impact of using biased data. The first challenge in assessing and analyzing these datasets lies in understanding how the data are collected and curated. Unfortunately, due to data privacy laws,[9] it is difficult to gather detailed information about these two steps. Consequently, this chapter focuses on the (lack of) accuracy rather than on the nature and consequences of potential biases. In sum, the overall objective of this research is to develop a methodology and a narrative—which in this case is to some degree specific to the aforementioned datasets—for data collection and curation. However, I hope this approach serves as a starting point for further discussions on data quality and availability in air transport research.

[9] https://www.bts.gov/confidentiality

Chapter 4: Spatio-temporal dynamics in airport catchment areas—the case of the New York Multi-Airport Region

Filipe Teixeira	Conceptualization, formal analysis, investigation, methodology, software development, visualization, writing
Ben Derudder	Conceptualization; article draft, review, and editing

Using the example of domestic connections departing from the New York Metropolitan Area, this chapter contributes to research on airports' catchment areas in MARs by exploring their spatio-temporal dynamics. From a data perspective, MARs offer a perfect setting to build an example of the challenges faced when analyzing large amounts of heterogenous data (e.g., multiple data sources). In this case, as I am looking into the underlying dynamics of airport attractiveness in MARs, the data used in the analysis span an entire year, with four daily observations, three times a week. One of the possible reasons why MARs research has not considered the dynamic nature of catchment areas relates to the challenges in analyzing large datasets (e.g., computation power) as well as due to the lack of adequate data and its vast heterogeneity. However, the past few years have seen the emergence of new data sources (e.g., HERE maps) and the consolidation of others (e.g., DB1B, T-100, On-Time performance). While examples such as HERE[10] maps have been fueled by the Internet of Things[11] (IoT) movement, historically established databases such as the BTS supported datasets have been given a new life through the emergence of APIs or statistical packages (e.g., SKYNET) that allow for easier data manipulation (e.g., collection, parsing, analysis).

Given that previous research has consistently shown that airport accessibility and different elements of airport utility (fare, connectivity characteristics, on-time performance) are key drivers of airport choice, I draw on the analogy with Huff models to calculate airport attractiveness to passengers in different census block groups. I marshal data sources that allow for an assessment of the spatio-temporal variability in the accessibility and utility of airports, which allows comparing catchment areas for different times of the day, days of the week, and quarters of the year, and for different utilities as well as overall utility.

Fueled by a new wave of data and data-driven tools (e.g., SKYNET), this chapter aims to shed light on the different types of dynamics in MARs. First, this approach differs from earlier research on airport choice in MARs, which was largely built on revealed or stated preference approaches rooted in the use of survey data. The motivation for this new approach is obviously fueled by data availability as well as by some of the shortcomings of using survey data. MARs are context-dependent and surveys are therefore to some degree idiosyncratic, which makes comparing and generalizing across MARs difficult (cf. Fuellhart & O'Connor, 2019). With this in mind, our results reveal different types of dynamics, and can be used as the input to follow-up research. We argue that such a model-based approach holds major potential in comparative research and/or research on MAR dynamics, but should be fine-tuned through the use of other data sources and/or refined model specifications.

[10] HERE maps provides historical and real-time traffic data through the ArcGIS platform (https://www.here.com)
[11] System of interrelated and interconnected computing devices with the ability of transferring data over the internet.

Chapter 5: Visualizing the potential for transit-oriented development—the development of an open and interactive planning support tool in Flanders, Belgium

Filipe Teixeira*	Conceptualization; data curation; software development; visualization; article review and editing
Freke Caset*	Conceptualization; formal analysis; investigation; methodology; writing

*Ghent University, Belgium.

In this chapter, I present StationsRadar, a data-driven web-based tool that was developed to support integrated land use and transport strategy making at railway station locations. StationsRadar was first developed as a way of transferring data between my colleague Freke Caset and myself. Initially this tool existed in the form of R and Shiny, as there was limited interaction needed (e.g., plot radar diagrams). As my colleague's project progressed, the need for more complex visualizations started to emerge. For example, there was the need to include several layers of information with the plotted maps. As data grew in size and complexity and as more functionalities were added to StationsRadar, it became evident that this tool's potential went beyond transferring data between two researchers. With this in mind, StationsRadar grew into a data-driven web-based tool, aimed at a larger audience as described below. We set our geographical focus on the region of Flanders and Brussels Capital Region. This tool classifies as an "accessibility instrument" (Silva et al. 2019) as it communicates the empirical findings that resulted from a systematic appraisal of the accessibility of railway station locations in both regions. We developed the tool in close dialogue with policy and planning stakeholders by drawing on the experiential case study research strategy as recently proposed for planning research by Straatemeier et al. (2010) (see also Straatemeier 2019 and the many contributions discussed in Silva et al. 2019). In doing so, the chapter echoes the widely shared contention within current debates on planning support systems (PSSs) (and on accessibility instruments in particular) that, instead of developing ever more advanced tools, more research is needed that closely examines actual user experiences and expectations, and that also explicitly considers the local planning and institutional context (Silva et al. 2017 and 2019, Silva and Larsson 2018).

Chapter 6: Discussion and final remarks

The final chapter of this dissertation first describes the state-of-the-art and summary of findings (6.1), followed by the limitations of my research and, associated with this, possible future avenues of research (6.2). I conclude with some final remarks and reflections on the potential of data-driven tools to reduce data complexity and enhance data democratization (6.3).

1.2 Availability and usage of data in transport geography

The number of research articles focusing on data-driven tools increased from less than 250 per year in 2010 to nearly 2,000 in 2019 (Figure 1). While a thorough bibliometric analysis would be needed to better comprehend the processes underlying this fast growth, it can be assumed that Figure 1 is in large part representative of the growing interest in data-driven tools. Clearly, this is in part fueled by the ubiquitous stream of data and the proliferation of computer languages that have opened the access to better understanding the data under scrutiny (Miller & Goodchild, 2015). In the remainder of this chapter I will discuss the impact of data on theory and on research. I will also explore whether data has the ability to drive new research methodologies and tools.

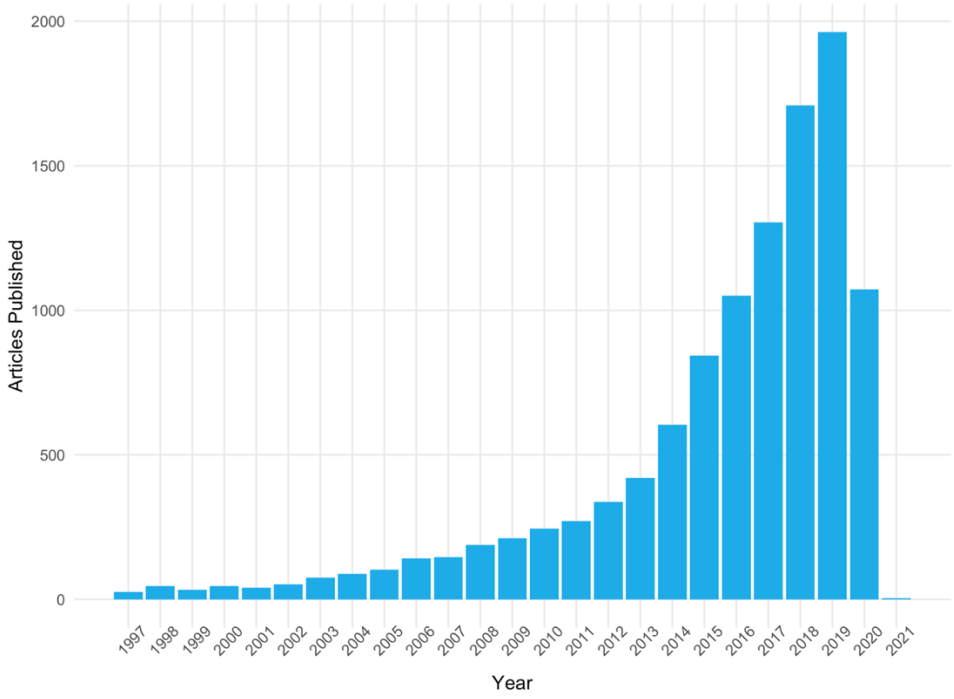


Figure 1—Research articles mentioning "data-driven" in their title. Extracted from Web of Science (11/08/20).

Today it is difficult to overlook the impact on our daily lives of a world that is constantly connected to the "world wide web." For the past decade, the process of collecting, storing, and processing data moved from large specialized data centers to machine learning-powered devices that fit inside our pocket. When the iPod was first introduced in 2001, it held a capacity of 5Gb and had only one functionality: playing music. Today, iPhones have a capacity that is over tenfold and include built-in Artificial Intelligence (AI). This increased capacity in computing allows every person owning a smartphone or smart device to become a beacon of data. These fundamental changes in data collection are in part responsible for the birth of the era of Big Data.

These large volumes of data are clearly a source of enthusiasm in science. In his infamous *WIRED* magazine article, Chris Anderson (2008) proposed the idea that Big Data could spell the end of theory in science. While this piece was heavily criticized for its naivety, it raised an important question: What is the meaning of Big Data in research and how does it change our (i.e., researchers') position, methodologies, and epistemologies towards facts? It is unreasonable to conclude that Big Data would

mean the end of theory, as not even the most advanced artificial intelligence is able to execute human-like cognitive tasks (e.g., answer a question without first being taught how to answer it). However, despite data not being itself enough to answer most research questions, this has not stopped the enthusiasm for publishing data-driven research. Kosinski et al. (2013) published what would pave the path to one of the most controversial cases involving the use of social media data to predict human traits and subsequently influence behavior (Isaak & Hanna, 2018; Ward, 2018). While Kosinski's paper is by no means a hallmark of Big Data research, it does show a trend of using Big Data in research.

This flood of observable data has created new opportunities in the social sciences, which until now have often been operating in so-called data deserts (Kitchin, 2013). Historically, the social sciences have depended on data derived from surveys or national censuses, where the information is often limited, infrequent, geographically pre-organized (e.g., by census tracts), and restricted. This trend, however, has been reversed in the past decade, in part supported by the consolidation of GIS (e.g., ArcGIS,[12] QGIS[13]) and open data initiatives (e.g., OSM, BTS datasets). However, before exploring the relationship between Big Data and geography as a scientific discipline, it is important to ask when data becomes "big." Contemporary data became extraordinarily large only recently. For example, in 1969 the Apollo 11 program relied on an onboard computer with 4KB of RAM,[14] supplemented by 5 IBM[15] computers on the ground, to make the calculations necessary to land the lunar module. Today, even the most basic smartphone has about 4GB of RAM, which is about 1,000,000 times of what was used in aerospatial engineering 51 years ago. The term Big Data itself only emerged in the 1990s (Lohr, 2012, 2013, 2014). With this in mind, we can argue that data is "big" according to contemporary notions of how we are able to handle it. The size or volume of what we understand as Big Data will most likely be considered insignificant even years from now. However, in this dissertation, the concept of Big Data serves as a starting point to understanding the challenges of working with data. Big Data's existence as a concept has led researchers to develop structured approaches to how to approach it and deal with its challenges. Therefore, it sets the scene for approaching any other type of data that is deemed "large" by those handling it.

1.2.1. Big Data: Definition and challenges in social sciences

Despite the often-gratuitous use of the term "Big Data" by news media outlets to coin any large dataset, data scientists have agreed on a straightforward definition to characterize this type of data. Big Data comprises datasets that are characterized by three dimensions (Kitchin, 2013; Miller & Goodchild, 2015): (1) volume—the amount of data that can be collected and stored; (2) velocity—data are often created in or close to real time; and (3) variety—the diversity of structured data (e.g., organized and stored in tables and relations) and unstructured data (e.g., text, images). Alongside these three dimensions, another element that characterizes Big Data are the sources it originates from. Big Data sources can be divided into three categories: (1) directed—by means of digital surveillance; (2) automated—where a system (e.g., smartphone) collects data without the direct intervention of the user; and (3) volunteered—where users (i.e., often with IT technical skills) feed data into a common system such as OpenStreetMaps[16] (OSM) or other similar systems (Kitchin, 2013).

[12] https://www.arcgis.com/index.html
[13] https://www.qgis.org/en/site/
[14] Random-Access Memory
[15] https://www.ibm.com/ibm/history/ibm100/us/en/icons/apollo/breakthroughs/
[16] https://www.openstreetmap.org/about

Despite the long history of geography's relationship with large datasets, some argue that beyond GIS science other fields of geography have been slow to adjust to the data revolution (Floridi, 2012; Graham & Shelton, 2013). Kitchin (2013) argues that human geography has been rather unprepared for this data revolution, with only a few research centers and scholars being up to date. Interestingly, for the past few decades we have been witnessing researchers from physics to computer and data science venturing to make assertions about social and spatial sciences. Partially fueled and backed up by a set of skills (e.g., computer programming, modelling, and simulation) often innate to fields such as physics, the field of social physics has been contributing to the social sciences. However, despite the momentum, these new emerging fields often ignore some important traditions in urban quantitative analysis and model building (Bettencourt et al., 2007; Jonah, 2010; O'Sullivan & Manson, 2015). As a result, there is often a naïve and disconnected view on cities that fails to consider elements otherwise fundamental in the social sciences (e.g., politics, culture, policy, capital).

1.2.2. Open data

Another challenge pertains to the availability of data. Despite the flood of data, it is paradoxical that most of the access to such data is limited. The reason behind such limitations is that most data has been generated and collected by private business, with a few exceptions being generated by governments (e.g., BTS datasets). However, interest in "open data" has been rising during the past few years, with a few notable examples such as the EU Open Data Portal,[17] UN Databank,[18] Canada Open Government Portal,[19] and the New York City Open Data[20] initiatives. Despite the interest in such "open data" initiatives, commercial database providers have often been able to provide curated data, supported by a set of tools and features allowing an easier analysis of the data under scrutiny (e.g., OAG, Esri[21]) (Marques Teixeira & Derudder, 2020). These challenges in data accessibility often lead researchers to turn to social media (e.g., Twitter, Facebook, Instagram), as it offers a stream of ubiquitous multivariate data, with easy-to-access platforms (e.g., APIs, online tools) (Ruths & Pfeffer, 2014; Szell et al., 2014; Poorthuis & Zook, 2017). Although social media is an interesting source of data, the risks often seem to outweigh the advantages (Morstatter et al., 2014). One of the first challenges of using data originating from social media networks is to find how representative the data is. Following the example of Twitter, only a small fraction of the tweets is accurately geolocated (Poorthuis et al., 2014). Until 2015, tweets would include accurate GPS coordinates in their metadata (Drakonakis et al., 2019), exposing a user's location and allowing researchers to study individual movements. However, this feature would later be revoked and brought into voluntary participation only. This adds to the fact that despite social networks being able to represent populations rather than samples (Miller & Goodchild, 2015), these populations are still self-selected. In fact, despite the global use of social media, what could seem like populations are just large samples of people who, for example, signed up for Twitter or carry a smartphone. While social media can constitute an interesting source for geolocated data, it is important for researchers to scrutinize social media data for bias. However, some of these challenges should not herald the end of a data-driven geography. The lack or

[17] https://data.europa.eu/euodp/en/data/
[18] https://www.un.org/en/databases/
[19] https://open.canada.ca/en
[20] https://opendata.cityofnewyork.us
[21] https://www.esri.com/en-us/home

unevenness of data (e.g., from social networks) can, for example, be used to reveal patterns of inequality (e.g., access to smartphones or the Internet).

Alongside the widespread availability of spatial data, the following question emerges: does data-driven and data-centric science spell the end of theory? In the beginning of this section, we briefly mentioned Anderson's article (2008), in which he claimed that data abundance would mean the end of theory in science. While this view on data and its meaning to future science is still the target of criticism, for the past few years more researchers have been laying out the limitations of theory in the era of data-driven research. Rather than calling for the end of theory, Watts (2012) instead proposes a type of theory that addresses identifiable social phenomena instead of abstract entities such as the entire social system. This approach to theory follows Merton's (2007) call to middle-range theories (i.e., empirically grounded theories based on observations). Data science follows a similar approach by shifting from abstract and generalist theories towards the specific. Urban theory and planning have in the past focused on global or radical urban changes, showing little concern for smaller scale events (e.g., how local movements and small spaces sustained a city) (Batty, 2012).

However, as more data is made available, new patterns (e.g., shopping habits, short distance travel) start to emerge as researchers are able to make observations at a smaller scale than before (see Poorthuis et al., 2014; Poorthuis & Zook, 2015). Poorthuis and Zook (2014) mapped the distribution of selected cultural-economic indicators and self-defined identities (e.g., users that self-identify as bankers or artists) in the New York Metropolitan region by using georeferenced tweets. Szell and his colleagues used a data-driven web tool (Offenhuber et al., 2014; Santi et al., 2014) as a way to understand the linkages between travel habits and the places travelled to and from most often in the New York Metropolitan region. Neal (2014) used the BTS DB1B dataset to map seasonal differences as well as to understand the differences between business and leisure travelers in air travel in the US. The shift from macro scale analyses, where processes are often aggregated, to micro analyses, where researchers focus on individual processes (e.g., individual movements) in geography can to a certain extent be compared to evolution in the biological sciences. While microscopes had existed since the 16th century, it was not until the 19th century that we witnessed the birth of what is now called microbiology. Biologists shifted then from observing processes at a macroscale, to doing observations at a microscale, allowing researchers to answer previously unanswered questions.

While this flood of observable data is unlikely to be the answer to most of geography's questions, we can argue that it harbors the potential of being the starting point for a new stream of geographical research. By leaving aggregated data behind, researchers can focus on different types of dynamics. For example, MARs research has predominantly perceived these regions as static elements in space and time. Currently we have data that allow us to study, first, accessibility to the airport for different times of the day, week, and year, and second, individual flight and passenger movements. In this example, we are able to shift from a static concept using aggregated data to a concept that considers the spatio-temporal dynamics in these regions.

1.3 Challenges in acquiring, curating, analyzing, and visualizing data

In this section, I will explore the process and challenges in the four stages of working with data: data acquisition, curation, analysis, and visualization. Today, data is mostly provided through an API, which facilitates the access to data by defining a set of routines, protocols, and tools. While there is a vast number of different types of APIs available, we can loosely fit them into two main groups: full-service APIs (e.g., Google, Esri, Mapbox[22]) where users have access to a solid, diverse, fully matured set of datasets and tools, supported by different programming languages, and basic-service APIs (e.g., Flight Aware,[23] Flight Scanner,[24] Kayak,[25] Expedia[26]), often offering access to a single dataset, limited availability of built-in tools, and limited support (i.e., in terms of programming languages supported or technical support). Because full-service APIs are more consistent and offer a better overall service, they are often considerably more expensive than basic-service APIs. A good example of a commonly used full-service API is offered by the OAG. This data provider offers a suite of air transport data, backed by insightful analytics, tools, and a mature API with the possibility to export the data to different formats. However, the high price tag might not always be an option to most research groups with limited funding. With this in mind, most researchers often opt for an affordable alternative, despite the challenges it bears (e.g., no dedicated API, lack of technical support, format of exported data). In the case of air transport research, commonly used alternatives for US based research are the DB1B and T-100 datasets. Unfortunately, and as we further describe in Chapter 2, these free-to-use alternatives come at the cost of the BTS not having an API available to download the data. Actually, most BTS datasets are made available by means of individual CSV files, making a longitudinal analysis complex and cumbersome in most cases. With that in mind, I created SKYNET, which will be thoroughly described in Chapter 2. While SKYNET does not intend to be or to have the functionality of an API, it aims to remove some of the shortcomings from the absence of one (e.g., facilitating downloads, parsing data on demand).

1.3.1. Tools: Processing and storage

Over the past few years, the proliferation of programming languages such as R, Python,[27] and JavaScript[28] have supported the emergence of software that allows researchers to interact with data from providers lacking or with very limited APIs (e.g., SKYNET). This software often comes as a package in R or as a library in Python and JavaScript, and builds on the concept of web scraping. In contrast to an API, where there is a consistency in the rules and ways of accessing the dataset by the data provider, web-scraping extracts information from the provider without any formal procedure. There are several challenges associated with the web-scraping method, the first one being the lack of homogeneity in how the data is queried. For example, when using an API to access a data provider (e.g., database, website) the results of a query are homogenized to make sure that regardless of who accesses the API, they get similar results (e.g., variables, data structure). As web-scraping does not use any formal methods, the variables extracted or even the destination data format (e.g., CSV,

[22] https://www.mapbox.com
[23] http://flightaware.com
[24] https://flightscanner.com/en
[25] https://www.kayak.com/
[26] https://www.expedia.com/
[27] Interpreted, object-oriented, high-level, general-purpose language (https://www.python.org)
[28] High-level and multi-paradigm language (https://www.javascript.com)

JSON[29], SQL[30]) might differ from user to user. Fortunately, there have been some efforts to create a syntax of spatio-temporal data (i.e., starting by how it is accessed), mostly by R developers (e.g., Beautiful Soup,[31] Tidyverse's rvest[32]).

Regardless of the type of API or if the access to the data is unrestricted (e.g., open access) or behind a paywall, data needs to be curated following acquisition. Unfortunately, most data sources are often messy, requiring substantial processing and curation (Miller & Goodchild, 2015). Except for governmental data where there are often strict regulations on how to collect, parse, and curate data (e.g., Code of Federal Regulations in the US), data tends to come in an unstructured format, lacking documentation or metadata or having ambiguous quality control. For example, air transport data providers (e.g., RDC Aviation,[33] OAG) do not disclose information on how the data is collected and treated, invoking the argument of data privacy and competition laws. This leads to researchers being left with having to trust the data they use. However, as I will further explain in Chapter 3, even data following strict regulations (e.g., BTS datasets) raises concerns of being potentially biased. Unfortunately, in what I assume to be related to time costs and efforts necessary to analyze data, most commonly used datasets are rarely put under scrutiny.

1.3.2. Storage

With this in mind, data curation can have three important stages: storage, parsing, and validation. Depending on the size, characteristics, and final purpose of the data, researchers are faced with a plethora of data storage solutions. This often means that when choosing the best storage solution, researchers seldomly have as a priority an interchangeable format that can later be easily used by other research groups. While it is not the goal of this dissertation to provide an extensive list of the available storage solutions, an easy and probably more commonly used storage option is by means of CSV or JSON files. However, and despite the ease of transferring CSV or JSON files between different platforms, they are hardly scalable on their own or specific to spatio-temporal data. Some interesting options (e.g., geoJSON,[34] ElasticSearch[35]) have emerged over the past few years, facilitating the usage and transferability of spatio-temporal data. However, the absence of a formal arrangement of storage remains in geospatial sciences. I acknowledge that it is to a certain extent unrealistic to believe that different types of spatio-temporal research, and the focus of this dissertation (i.e., transport geography research), could benefit from a single data storage solution.

1.3.3. Curation

The absence of formal arrangements for data curation is only accentuated by the ambiguity of geographic concepts such as neighborhoods, regions, and developing countries, which can be vague, fluid, and contested (Miller & Goodchild, 2015). These concepts, like most geographical knowledge, are buried in theories and models, and use an informal language that must be adjusted to be computed

[29] JavaScript Object Notation
[30] Structured Query Language. Programming language used for managing data held in a relational database management system.
[31] https://www.crummy.com/software/BeautifulSoup/
[32] https://rvest.tidyverse.org
[33] https://www.rdcaviation.com
[34] Open standard format, based on JSON, designed to represent spatial features (https://geojson.org)
[35] https://www.elastic.co

and readable by machines. This creates a need to formalize these concepts into a machine-readable format.

To avoid some of the challenges in data curation, the first step is to ensure that data is scalable, reproducible, and replicable (Patil et al., 2016). The focus on these three characteristics should ensure that different users supported by different platforms (e.g., programming languages, operating systems, GIS systems) can use data interchangeably. For example, both CSV and JSON are excellent solutions for data storage as they are platform-agnostic.[36] As they are not associated with a particular software (e.g., geodatabases and ArcGIS), they can be easily imported and manipulated by most programming languages. In terms of scalability, R has the data.table[37] package, which allows importing considerably larger CSV files (e.g., 500mb) in just a few seconds. On the other hand, ElasticSearch provides an alternative to NoSQL[38] databases. ElasticSearch provides a full-text search engine, with an HTTP web interface and schema-free[39] JSON documents. By using JSON files, this solution considerably decreases the effort often put into importing data into a database.

In Chapter 3, I elaborate further on the challenges of working with potentially biased data by focusing on the example of the BTS DB1B and T-100 datasets. However, the challenges of the BTS datasets can often be found in other datasets as well. For example, despite the data collection directives and regulations defined by the BTS and the FAA, we were able to identify potential biases in the DB1B dataset. Unfortunately, the data collection methods are unclear, leaving the window open for assumptions regarding what could have caused the bias. Cases where widely used datasets have potential biases despite being heavily regulated are common in data-driven science (Morstatter et al., 2014; Malik et al., 2015; Lum & Isaac, 2016). However, as the causes of bias are often difficult to ascertain, researchers could profit from developing systems that would allow the identification of bias in data. We can think of such systems as warning methods, which in turn would equip the researcher with the tools necessary to identify potentially biased data before further analysis and assertions are made.

Beyond storage, parsing, and validation, curated data often lack the configuration needed for integration (i.e., in order to associate and integrate with other data sources) and preservation (i.e., maintaining archived data to ensure it can be accessed through changes in technology) (Stonebraker et al., 2013). This leads to "single-use" data[40] being produced, regardless of the considerable efforts put into the previous stages of working with data. Despite not being a new concept, "single-use" data have arguably become more frequent, giving way to situations such as the replication crisis in psychology (Maxwell et al., 2015). This crisis in psychology was in part fueled by the lack of structure and standardized storage in the collected data. As psychological sciences often rely upon single case studies, interviews, or clinical studies, data often come in a non-standard form (e.g., often still stored in paper archives), making digitalization, integration, and posterior replication difficult (Notaerts et al., 2017). As most research groups seldomly have a data curation plan, data is used on an ad hoc basis,

[36] Not sensitive or fixed to a particular platform
[37] https://rdatatable.gitlab.io/data.table/
[38] 'Non-SQL'. Database which is modeled in means other than the tabular structure (e.g. SQL)
[39] Data can be stored without a previous structure
[40] Data processed, curated and stored with only one research project in mind.

and buried in hard-drives or personal computers, making it often inaccessible to future researchers working on the same topic.

While it may seem paradoxical, as the capacity to collect, store, and process data grows rapidly, the ability to analyze this data grows at a much slower pace (Keim et al., 2006). In the past, researchers working with large datasets often had to reduce data prior to analysis. One of the reasons is that most traditional statistical methods do not scale well with high-dimensional data. In the case of high-dimensional spatial data, some of the statistical methods still used were initially developed with other applications in mind, and lack specificity, which in turn brings other challenges. Some of the statistical methods applied to high-dimensional data were developed for the fields of genomics (e.g., genome sequencing, biochemical pathways), neuroscience (e.g., brain connectivity networks, fMRI data), and economics (e.g., risk management, stock analysis). This means that until large amounts of GPS or satellite data were generated in the context of geospatial sciences, most high-dimensional statistical techniques did not take spatial dimensions into account (Fan et al., 2014). Currently, even some of the most commonly used methods for data analysis (e.g., network analysis, clustering), are seldomly specific to spatial data. However, there have been some considerable advances in this domain, with some concepts of spatial clustering analysis (Entwisle et al., 1997; Jacquez, 2008) and spatial network analysis (Scheurer et al., 2009; Zhong et al., 2014) being used more often in the field of geospatial sciences. While some of these concepts are not new, they have been clearly propelled and made more accessible by technological advances (e.g., computing power, easier access to cluster computing, machine learning).

1.3.4. Visualizing data

The inconsistency in how we analyze data, together with the lack of specific tools (e.g., for spatio-temporal data), is leading to a crisis in which scientists collect and store more data than they can analyze (Floridi, 2012; Fan et al., 2014). In fact, most data analysis systems still rely on interaction metaphors[41] developed decades ago (Keim et al., 2006). Interaction metaphors have been widely used in several contexts, but they are mostly known through the context of Operating Systems (OS). When personal computers started to become widespread in the 1980s, their command-line only interface[42] did not allow an easy interaction with the OS unless the user had some affinity with IT systems. Later, some OSs started to include a Graphic User Interface[43] (GUI) (e.g., Macintosh Desktop, Windows 1.0). Instead of having users typing long streams of code to copy or access files, interaction metaphors, such as files and folders, where developed to simplify and conceptualize the process. The same principle is often used in the field of visual analytics.

The emerging field of visual analytics has been a promising solution for bridging the gap between collecting and analyzing data. This emerging field combines research fields such as visualization, data mining, and statistics, focusing on handling large, dynamic, and heterogenous data. The field of visual analytics uses interaction visualization techniques and algorithms alongside data analysis methods in order to support analytical reasoning for decision making. However, and despite the influences from a broad spectrum of scientific fields, the overarching goal of visual analytics is clear: to turn the data flood

[41] Set of user interface visuals, actions and procedures, that exploit specific knowledge that users already have of other domains

[42] A command-line interface processes commands to a computer program in form of lines of text.

[43] Form of user interface which allows users to interact with computer programs and electronics, through visual cues.

and information overload into an opportunity. In order to understand some of the challenges the field of visual analytics faces, we have to revisit the concept of data democratization. As I have mentioned before, one of the challenges of data democratization is to find ways of communicating complex data analytics to "non-technical" people while retaining the original information being conveyed.

In order to simplify interaction and data visualization, the field of visual analytics adopted the concept of interaction metaphors, which I introduced in the previous chapter. Visual analytics scientists use interaction and visual metaphors to, for example, make high-dimensional or complex data accessible to everyone (Andrienko et al., 2010). This need of visual analytics to aid both data analysis and visualization was one of the core reasons behind StationsRadar, which I will explain in Chapter 4. StationsRadar was to a certain extent created from the interaction between data analysis and visualization. This interaction can be linked to the field of visual analytics, as visualizations can be used for analysis, communication, and decision making. However, building visualizations that can be used for different domains and purposes harbors challenges as it is not always a straightforward task. With this in mind I will continue discussing the challenges in data analysis, gradually introducing the step of data visualization. While the challenges of visual analytics are mostly field-dependent (e.g., economics data generated every second, bioinformatics and genome data with billions of base pairs[44]), there are four identifiable core challenges that are common to data analysis: provenance, semantics, user acceptability, and scalability (Keim et al., 2006; Keim Daniel & Mansmann, 2010). Data provenance is the science of understanding the origins of data, how it arrived in the user's database and how it was curated (Buneman et al., 2001). In the field of biochemistry, and more specifically, with genome-wide association studies[45] (GWA), data is often made available on the web, copied to other databases and curated several times (Keim et al., 2008). However, the lack of metadata and documentation makes it difficult to reproduce results and combine multiple data sources. The absence of metadata and documentation yields challenges similar to the ones already mentioned for data curation. Provenance in scientific data focuses on three dimensions: data provenance, the source of data, and the link between source and the system using it; analytical provenance (i.e., curation), processes performed on the data; and reasoning provenance, how and why analysts arrive at their conclusions. By understanding the sources and the transformations applied to the data, it is easier to explore the links between evidence and hypothesis. This link will in turn allow a better validation, scalability, and manipulation of data as it increases its transparency (e.g., higher data quality results in an easier to analyze data) (Varga & Varga, 2016).

In visual analytics, semantics refers to the act of extracting meaning from data. This is very likely the most complex topic as, first, it is heavily connected to user acceptability, and second, the more stakeholders a dataset targets, the broader the derived semantics will be. In transport geography and more specifically in the case of StationsRadar, it is not uncommon to have stakeholders (e.g., policy makers, urban planners) with different needs and knowledge (i.e., in terms of data and how it is presented). In the case of StationsRadar, when deriving semantics from data, we had to consider not just policy makers and urban planning stakeholders, but also different needs on the regional scale (e.g., different rail corridors). A good example of semantics in data is what some refer to as the Web 3.0 or Semantic Web (Berners-Lee et al., 2001). The semantic web is an extension of the World Wide Web

[44] Fundamental unit of double-stranded nucleic acids. They form the building blocks of the DNA double helix.
[45] Observational study that collects DNA of a different participants and compares their entire genome in order to find commonalities in a particular trait or disease.

through a series of projects and standards that make Internet data more machine-readable. By using metadata to encode semantics (e.g., reasoning, knowledge) into data, it is possible to establish relationships between entities and categories of things, and with that, establish automated reasoning using data or facilitate operating with high-dimensional or heterogenous data. Encoding data with semantics does not refer exclusively to the addition of keywords to the metadata. For example, visuals, color palettes, or even pictograms can be another form of representing a concept or knowledge extracted from data. As visuals are built both for analysis and data visualization, it is important to ensure when reducing or aggregating the data that no information is overlooked. With this in mind, it was important to use intuitive graphics, agnostic to the field of knowledge of the stakeholder, which in turn could complement the displayed maps in our tool. Besides choosing graphs, maps, and how to present data in a format acceptable by users, it is important to consider the design elements to use (e.g., color palette, website design, text fonts). StationsRadar uses a color palette (Figure 2) that considers different design elements (e.g., types of screen, color-blindness, print-compatible).

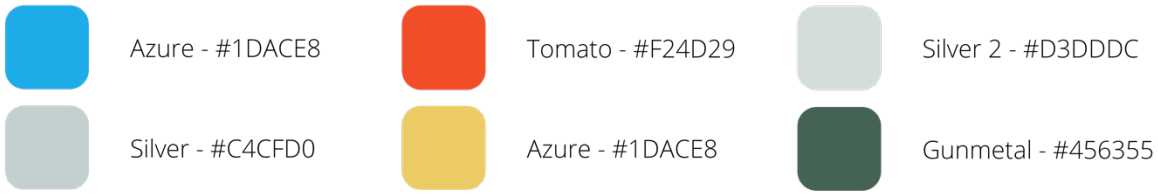


Figure 2—Color palette used throughout StationsRadar.

1.3.5. Scalable data

Finally, we have scalability. When developing visual analytics, it is important to ensure that they are scalable (e.g., shape, size). For example, in Citizen Science,[46] where different users with different backgrounds (i.e., mostly amateur scientists) contribute to and participate in research, it is important to understand that the visual analytics consider an increase in participation (e.g., more users adding or analyzing data) or dimensional changes (e.g., more variables/dimensions being included). In order to avoid any setbacks, it is important to foresee how the dataset will increase in size and shape, and how these two elements will impact both visuals (e.g., graphs, maps) and tool design (e.g., menus, text). For example, in StationsRadar we use radar diagrams to display data. However, we acknowledge that this type of diagram does not scale well if we have to add extra dimensions or variables, as radar plots can only visually bear a limited number of dimensions. In the case of StationsRadar this is unlikely to become an issue as we do not foresee adding more dimensions or variables. The second challenge in scalability is in assuring that results of data analysis and visual analytics are readable. For example, if a tool (e.g., SKYNET, StationsRadar) is to have a redesign or iteration in order to follow advances in technology, it is important to ensure that programming code, visuals, and derived semantics are consistent, readable, and clear. A solution for this challenge is to develop a transferable syntax. In computer science, a syntax is comparable to the concept of grammar in languages. This syntax or grammar has been more common in visual analytics (e.g., grammar of graphics, Vega visualization grammar).

[46] Scientific research conducted by amateur scientists

1.4 Thinking critically about a data tool-based geography

In order to critically reflect on the position of research that is bent towards the development of analytical tools within geography at large, it is useful to briefly revisit earlier debates within geography on the position of GIS. Revisiting these debates is useful because it allows me to clarify my own position on how I believe my dissertation sits within the discipline at large.

The emergence of GIS in the 1950s and its widespread availability in the 1980s paved the way for a revival of quantitative geography that could play a central role in the social sciences. Openshaw (1991) approached these shifts with optimism and even with enthusiasm. It is clear from his commentary in *Environment and Planning A* that he takes a positivist view on what the GIS could mean for the future of geography. Following this techno-positivist approach, Openshaw argues that as computing power becomes more accessible and more mainstream, science will be driven to an "immensely data-rich but theory-poor world." Some of the arguments presented by techno-positivist social scientists such as Openshaw (1991) revolve mostly around excessively theoretical approaches found in less quantitative streams of science. Social scientists who are either agnostic to or simply do not believe in the value of GIS are portrayed as being technophobic, pseudo-philosophical, and against change in the social sciences. Openshaw refers to this movement resisting computer-powered social sciences as "building up a range of conceptio-theoretical arguments against it, express them in pseudophilosophical languages to provide a veneer of academic respectability, and a few misquotes from famous dead people who lived in a totally different world, and wait five years for the reaction to go critical" (Openshaw, 1991, p. 622). While this statement is surely uses hyperbole to make a clear stand against social scientists refusing to endorse the change brought by GIS, the statement is in my opinion still valid 30 years after it was first published. In sum, despite the heavily exaggerated language and tone of his comments, Openshaw clearly saw much potential in GIS as it delivers an "ad hoc" platform for "doing geography" exactly because, in contrast to other geographic epistemologies, it is data- and computer-based.

However, despite the enthusiasm from more positivist and quantitatively inclined geographers, disagreement quickly emerged that warned against the perils of a data-centric geography. Some researchers feared that those who were able to control data and maps would also be able to control the truth in a dystopian fashion. The fears extended to the belief that by focusing exclusively on data and tools to make "the most" of these, some of geography's other epistemologies would fade into insignificance. Smith (1992) argued that the new wave of tools and data in social sciences and more specifically in geography amounted to "exuberant" and "extravagant" approaches from "non-geographers." While his statement was hyperbolic, it had the merit of triggering further debates on the impacts of tool- and data-driven science. Smith (1992) argues that Openshaw's (1991) paper takes a naïve approach to GIS, ignoring decades of theory and debates in geography. He builds on the argument later supported by Schuurman (2000) that there is a clear difference between scientists and social scientists, with the former blindly following data.

There are two distinct elements to this discussion. The first revolves around the fears that exclusively relying on data and tools may lead those using it ignoring insights gathered in the past. To a certain extent these fears of an exclusively data-centric approach revolve around the many elements of geography and spatial sciences that cannot be quantified. However, the discussion on "how much can be measured" or "how much can be modelled" is not new to science, yet ongoing in the field of social

sciences. When cognitive behavioral therapy (CBT) was first developed in the 1960s, it represented an experimental-centric approach when compared to the early Lacanian psychoanalysis, which revolved strongly around Hegelian philosophy (Beck, 1993). The distrust of more quantitative methodologies and data, however, is not posited as a fear of the exact sciences or of mathematical approaches to human spatial interactions. Taylor (1990) argued that while GIS is well equipped to handle information, it fails to produce knowledge as it focuses on facts and is incapable of meaningful analysis. In sum, the movement of researchers distrusting how GIS was being positioned in geography, and subsequently geography, within the social sciences revolved around fears that by looking exclusively at data we would lose decades of epistemological reflections and theory-based research.

The second element of this discussion pertains to the impact of GIS and data-driven science as tools to convey truth and their overall impact in research and society. This "man in the high castle" (Dick, 1982; Freedman, 2013) view assumes that by controlling data and by controlling the way they are displayed, truth can seldomly be objective. GIS has been crucial in legal decisions that have resulted in millions of dollars in damages being paid to affected residents (e.g., Kennedy v. City of Zenesville, see Monger, 2010), or even in documenting systematic patterns of spatial inequity (Thatcher et al., 2015). In recent years, for example, we have been witnessing the emergence of citizen science (Silvertown, 2009; Townsend, 2014; Engin et al., 2020), which gives the control of the data back to those generating it. This is a clear departure from Smith's (1992) view on GIS-based geography, as the difference between social scientists, scientists, and citizens is muddled in the narrative that, by giving the data back to non-technical people, Big Data has the potential to expand scientific knowledge and increase scientific literacy (Bonney et al., 2009, 2014).

In response to this, the concept of "critical GIS" emerged. First introduced by Schuurman (2000), it focused on the impact of GIS technologies on people. The belief and core of concept of "critical GIS" is that GIS tools and data-driven social sciences harbor the potential to cause positive societal changes as well. Critical GIS offers a more constructive and balanced discussion. Taylor (1990), despite believing that GIS was in danger of transforming geography into a "trivial pursuit" science based exclusively on facts, argued as well that the potential of this new toolkit of computer-based approaches could have a positive impact when considered carefully. Goodchild (1991) furthermore argued that data-centric tools could lead geographers to question the databases and processes they use in research. This constructive discussion on the perils and benefits of a stream of social scientists supported by GIS tools led to most of what we know in "critical GIS." The concept of "critical GIS" stemmed from the idea that more spatial data, and more tools to analyze it and map it, would bring new possibilities alongside with new challenges in the field of geography (Schuurman, 2000). Thatcher et al. (2015) reflect on the possibilities of GIS and on how it has been used to reinforce or challenge social injustices. They do so by arguing that GIS is not merely a tool or a set of techniques, but rather a set of concepts that has been used to ask serious theoretical and empirical questions. For example, Thatcher refers to the conceptual notion of "justice," or passive and active conceptions of equality. Regardless of the concepts or numerous examples, the essence of "critical GIS" revolves around the usage of both mathematical and theorical concepts, alongside data-driven tools as a way of thinking critically about social dynamics. However, the core concept of "critical" itself can be criticized, as, in the case of the social sciences, it is represented through different streams and approaches (i.e., quantitative, qualitative). Regardless of the meaning of "critical," the essence of the "critical GIS" movement has over the past years led to constructive engagements between Geographic information science and its many derivations on the one hand and critical geography on the other hand. Rather than polarizing

opinions, "critical GIS" aims to ease the "science wars" by framing both perspectives as being valid in their own place.

In my view, the discussion on "who controls data controls the truth" should pay more attention to efforts discussing how data and science in general should be made more transparent and accessible to non-technical people. Importantly, by "non-technical people" I do not refer exclusively to non-scientists or non-quantitative scientists, but to anyone without the technical skills to handle and analyze data. To a certain extent, the idea of increasing transparency in data-driven science should be at the helm of discussions around the implications of data in spatial sciences. I argue that increasing transparency bears not just the potential to increase the trust in data-driven science and data, but to stimulate critical thinking by those disagree who are less technically capable.

Finally, when reflecting on the discussions presented in this section and on the opportunities and challenges associated with data-centric research, there is little to convince me of the usefulness of the "qualitative vs quantitative" war. While it is important to discuss the impact of data-centric conclusions and the position of theory in social sciences, it is not the discussion per se that I distrust but rather the segmentation whereas one body of scientists fights for the right of data and computer-based approaches, while the other states that theory is all. Instead, in my view the discussion should revolve around, first, how to make science accessible as we scholars run the risk of being perceived as pseudophilosophical and inaccessible (Hayes, 1992; Porter, 2009). Second, the energy put into these discussions should not revolve around who approaches the ground truth better (Openshaw, 1997), but instead on how can we make concessions from both sides in order to build better conclusions. With this in mind, in this dissertation I position myself as being in favor of a technocentric stream of research. Not because I believe that data-driven science is more valuable, but because I believe that technology allows us humans to answer questions we were not able to answer before.

1.5 Sub-objectives

In summary, the following chapters (i.e., Chapter 2 to Chapter 5), focus on the four stages of working with data (Figure 3). Chapter 2:, which focuses on data collection, introduces SKYNET, a flexible R package that allows generating bespoke air transport statistics for urban studies based on publicly available data from the BTS in the US. Chapter 3: focuses on data curation by exploring how potential biases in air transport datasets can be revealed and detailed by scrutinizing the BTS DB1B dataset. In Chapter 4: I explore the challenges in data analysis by investigating the dynamics of MARs by focusing on the New York MAR as a case study. Finally, Chapter 5: presents a data visualization tool (i.e., StationsRadar). A data-driven web-based tool, it is developed to support integrated land use and transport strategy making at railway station locations.

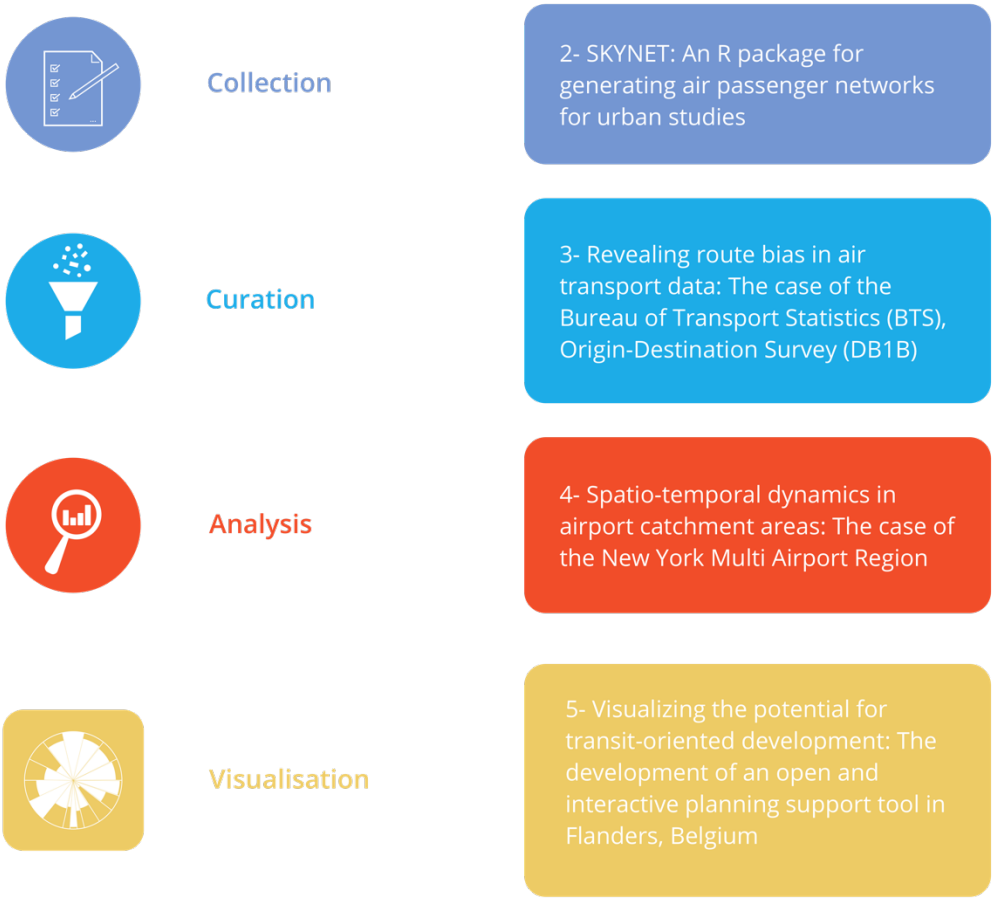
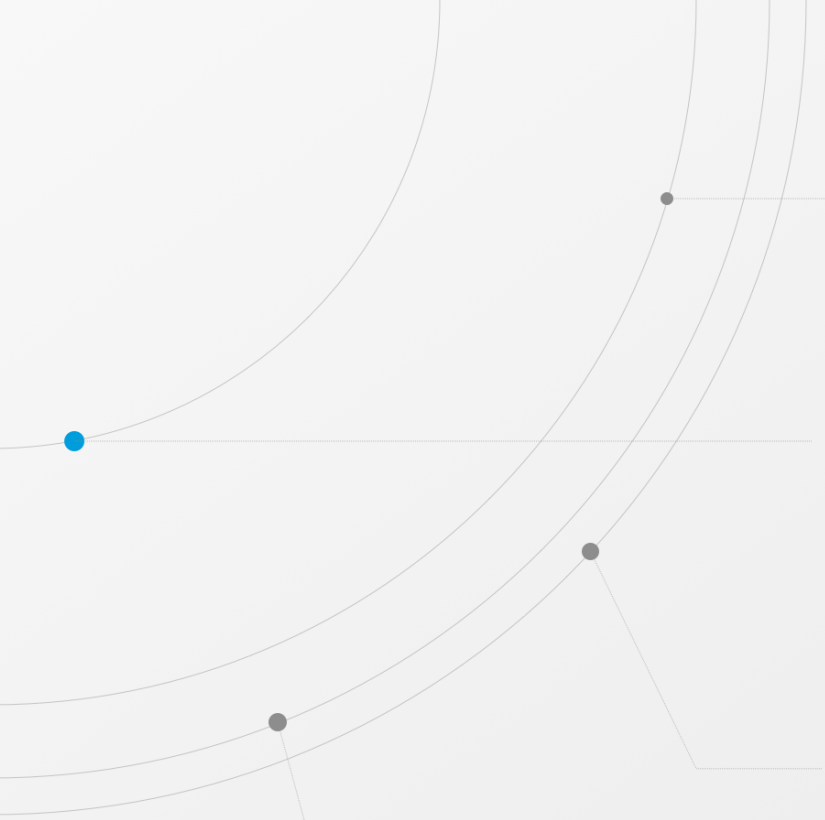


Figure 3 – Outline of the dissertation.



The Future of Networks Conference in Ghent. This was remarkably the first conference I have attended as a Geographer. It was undoubtedly one of the most interesting first contacts with Geography, mostly due to the people present in the conference.

I start developing SKYNET. This was the first real contact I had with R and most importantly with software development.

Time Geography Winter School Göteburg - Part II

Time Geography Winter School Linköping - Part I
First course in social sciences.

CHAPTER 2

Skynet – an R package for generating air passenger networks for urban studies

2.1 Introduction

Since Taaffe (1956) explored the relationship between the development of air transport in the United States on the one hand and its urban system on the other hand, there has been a blossoming and increasingly diverse urban studies literature looking at air transport-related issues. Research on the city/air transport-nexus ranges from substantive to more indirect analyses of that nexus. Examples of explicit city/air transport intersections are the analysis of Button and Yuan (2013) of the potential role that airfreight transport can play in stimulating urban-economic development and Goetz's (2000) analysis of the uneven geographies of urban accessibility in the United States after the onset of market deregulation in 1978. Examples of more indirect analyses often take the form of using airline statistics to make a broader argument about cities, such as Neal's (2014b) devising of a typology of hub cities by drawing on their role in air transport networks and Smith and Timberlake's (2001) analysis of the global urban system through the lens of air transport networks. Irrespective, urban scholars often have to manage with data that does not come in a format that allows for their immediate scrutiny in light of the research question at hand. This has prompted papers in the pages of this journal on air transport data issues in general (e.g. Derudder and Witlox, 2005) as well as on software tools to streamline the production of suitable statistics in particular (e.g. Neal, 2014a). In the latter paper, Neal presents a program – AIRNET – generating different types of intercity networks from publicly available data from the US Bureau of Transportation Statistics (BTS). In this paper, we present an enhanced programme – SKYNET – that (1) expands the options provided in AIRNET and (2) makes it more flexible in terms of its development and linking up with ancillary analytical tools.

The expansion resides in the fact that AIRNET was confined to generating different types of air transport networks. This is an important application but does not exhaust the data needs in research at the intersection of urban studies and air transport. The flexibility, in turn, emanates from SKYNET being an open source package in the R programming ecosystem. This allows researchers to have access to a major set of methodological tools without having to constantly transfer data between systems, as well as facilitating integration with major database solutions, network analysis and visualization tools, etc. In this methodology-focused paper, we summarize the main features of SKYNET, and show its practical use by zooming in on two research topics: (1) the literature on evolving urban landscapes of air travel accessibility (cf. Grubesic and Zook, 2007), and (2) the literature on the differences in intercity air transport networks by scale, type, and season (cf. Neal, 2014b). Throughout this paper the focus will mostly be on the United States, given that we draw on the publicly available data from the US Bureau of Transport Statistics (BTS). However, as with the selection of the two research topics, this is mainly for illustrative reasons: as we will show in the final section of this paper, SKYNET can be flexibly expanded to import and subsequently use other data sources in a similar vein.

The remainder of this methodological paper is organized as follows. First, we briefly review the two literatures we selected to showcase what SKYNET can bring to the table. Our purpose here is not to provide an exhaustive discussion of both literatures, but rather to point to the complex data (format) needs when tackling research questions that are commonly raised in these literatures. Second, we discuss some of the main features of SKYNET: the data it uses, the transformations it allows, and its potential to be flexibly expanded with related tools. Third, we show the usefulness of SKYNET by applying it to derive data that would allow researchers to efficiently address research questions in the two chosen literatures. And fourth and finally, the paper is concluded with a discussion of the avenues for further research.

2.2 Example literatures

2.2.1. Evolving urban landscapes of air travel accessibility

One of the biggest shifts in the urban geographies of air travel in the US was undoubtedly the national deregulation of the airline industry in 1978 (Gao, 1997; Goetz and Sutton, 1997). The concept of a free market in the US airline industry was introduced as a vehicle to enable decreasing airfares and increase the number of connections available to passengers. Although in general airfares did drop and connectivity did rise, a number of problematic side-effects emerged including the inter-related emergence of high-fare 'pockets of pain' and monopolistic behaviour of leading carriers at their hubs (Goetz, 2002). This led researchers to focus on issues such as the uneven urban geographies of nodal accessibility, service pricing and (changes in) competition (Goetz & Sutton, 1997; O'Kelly, 1998; Vowles, 2006). Consequently, and in parallel with large increases in available data, we have seen a growing number of research questions and methodological complexity in this literature (O'Kelly, 2016).

After deregulation, hub-and-spoke forms of organization emerged following the need to increase efficiency, reduce costs and exploit economies of scale and scope (O'Kelly & Miller, 1994; Goetz & Sutton, 1997; Goetz, 2002). The adoption of such systems emanated from the costs of adding additional spokes to a hub being relatively small in comparison to the benefits of having a more extensive range and capacity to their network. With further increases in passenger volumes came the need of improving the hub-and-spoke design (O'Kelly et al., 1996; Bania et al., 1998), as well as the widespread adoption of computerized reservation systems to increase the marketing reach and optimize the urban markets served (Levine, 1987; Mainzer, 2007).

However, the adoption of these systems has led to monopolistic behaviour through the establishment of 'fortress-hubs', nodes where a single airline controls a majority of the market and which enables it to set prices. While most larger carriers started adopting a hub-and-spoke organization, smaller carriers such as Alaska Airlines and Southwest Airlines adopted a more diffuse point-to-point form of organization (Bania et al., 1998). Although consequences have been complex and multifarious, many smaller communities have seen a decrease in service through schedule reductions, higher fares and elimination of routes. More than 25 studies over 20 years led the US Department of Transport (DOT) to identify the existence of 'pockets of pain' across the urban landscape. Goetz and Vowles (2000), for example, identified a cluster of cities with consistently high average fares and yields. This effect has been repeatedly associated with the lack of service from low-cost airlines and the dominance of some of the well-established and higher cost competitors, which resulted in an increase of fares without an associated increase in service (Goetz, 1993; Goetz & Sutton, 1997; Goetz & Vowles, 2000). It is clear that the urban-geographical effects of the market deregulation continue to unfold 40 years after its implementation. Or, as Wei and Grubesic (2016) recently put it: 'the pain persists'. Its continued effect on locational accessibility and the concomitant impact on urban economies makes the dynamics of the uneven pricing and connectivity landscape across the US urban system into a relevant research topic, albeit a research topic that often faces methodological constraints because of the complexity associated with handling the available data.

2.2.2. Mapping air traffic networks

Mapping and analysing (shifting) inter-city air traffic flows at the global, national and regional scales has been another point of attention in urban studies (e.g. Bagler, 2008; Guimera et al., 2005; Xu and Harriss, 2008). Several studies have focused on showing the importance of such networks (Borgatti

et al., 2009; Neal, 2014b), as well as analysing the different types of networks, their scalar dimensions, and their relevance in the context of different kinds of social and economic processes.

Inter-city air transport networks have been predominantly studied in light of their topological characteristics, focusing on three features in particular: (1) their small-world structure, characterized by the mean geodesic distance between nodes slowly increasing as a function of the number of nodes in the network; (2) their scale-free properties, for networks where the distribution of nodal degree follows a power law; and (3) their modularity, used by researchers to measure the strength of the division of the network into different sub-networks. However, past research has predominantly focused on the network of routes flown between airports and cities, neglecting other types of network characteristics. In an attempt to overturn this trend, Neal (2014b) compared some of the modular characteristics of the US air passenger network. By inspecting the weighted, dynamic and directed aspect of the US air traffic network on both substantive and methodological grounds, he develops a nested typology differentiating nodes' roles in terms of scale (airport vs metropolitan area), season (winter vs summer) and species (business vs leisure). For example, when inspecting US air transport network by season, Neal observes that only two-thirds of their edge sets overlap, as traffic increases during winter for the southernmost part of the US, and during summer for the northern half of the country. These structural differences are crucial when conducting research in the context of urban/air transport geographies, as caution should be exerted in selecting the appropriately defined network for the research questions at hand.

In the context of air transport research, seeing a network as set of Origin-Destination (OD) pairs, where passenger movements are shown, often brings an advantage over the traditional route networks which focus on airplane movements. The value of route networks when researching airline operations or logistics is undeniable, however passenger flows are more relevant to understand urban-economic dynamics (Button & Lall, 1999; Debbage & Delk, 2001; Neal, 2010) or city development (Derudder and Witlox, 2005; Smith and Timberlake, 2001) associated with air transport. Despite some criticism on the unclear lines that separate different types of travellers (Uriely, 2001; Lassen, 2006), the interest in understanding movements based on the passenger's requirements and characteristics has not faded over the years (Ostrowski et al., 1993; Chen, 2000; Neal, 2014b).

2.3 SKYNET

2.3.1. Data used

Many of the above-referenced papers make use of the provided by the United States Bureau of Transport Statistics (BTS). The popularity of this particular dataset can be traced back to the observation that airline data often tends to be expensive (Fuellhart et al., 2013) and/or lacking crucial information in light of the research question at hand (Derudder and Witlox, 2005). A major example of the latter is the lack of origin-destination data, which presents a major challenge when research questions target actual flows of passengers between cities rather than the set of airplane movements that constitute these flows. However, the BTS data – comprising the Origin and Destination Survey and the T-100 dataset, freely available from the BTS website[47] – allows circumventing many of these challenges this research field faces. In light of this, SKYNET makes use of the BTS data to show its functionality, albeit that the package is built so that other airline statistics can be easily imported after being transformed into a similar dataset structure. With that in mind, we have built into SKYNET the option of converting data from several major airline data providers (e.g. SABRE, OAG), into a format readable by the package. Both the vignettes and readme files provide all the necessary instructions regarding restrictions on variables and structure. However, for the sake of clarity, the practical examples discussed in this paper will zoom in on the case of the United States and the BTS data.

The Origin and Destination Survey (DB1B) comprises a 10% sample of airline tickets collected by the BTS. The data is organized into quarterly reports, starting in 1993 and updated quarterly, with the data being embargoed until the following quarter. It includes three different tables from which we will use the tickets ('DB1B Ticket') and coupons ('DB1B Coupon') data. A Coupon is an electronic or paper document or series of documents indicating the itinerary of a passenger. As shown in Figure 4, it reflects the discrete movements of a passenger. The Market variable reflects the uninterrupted (except for transfer) movements of a passenger. For example: An itinerary BOS-JFK-ATL-BOS with a break in ATL would have two markets, BOS-JFK-ATL and ATL-BOS. And finally, a Ticket shows the complete itinerary from start to finish, reflecting the complete movement of a passenger on a given airline ticket. Skynet uses the passenger variable as a relative frequency variable to size the nodes, which in this case reflect airports, and in the case of a weighted network as a tool to generate weighted edges.

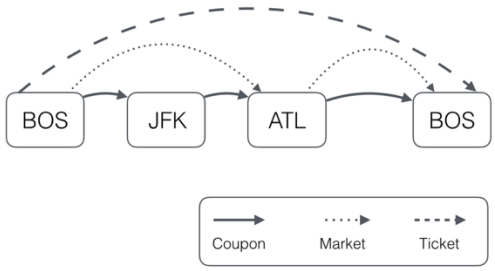


Figure 4 - Concept of 'trip' as for the BTS data.

[47] https://www.transtats.bts.gov/homepage.asp

Meanwhile, the T-100 Domestic segment dataset contains non-stop segment and market data reported by US and foreign carriers for Origin and Destination pairs located within the boundaries of the United States and its territories, organized on a monthly basis. Despite containing all flights for a given period, it does not include information that can be used to study passenger travel intentions or real passenger movements. However, as it includes all flight and passenger information, it becomes relevant when analyzing the entire network structure.

2.3.2. Skynet: Main features

Skynet was initially devised as an R package-based answer to some of the research challenges discussed in the previous section[48]. R was developed as a programming language for statistical computing and graphics, aimed to provide an array of tools for research in statistical methodologies. The need of having a tool seamlessly integrated into an environment capable of statistical and network analysis has arguably grown together with the volume of available data. For most US-related research questions, the BTS databases are perfectly capable to provide the researchers with enough raw data, and Skynet package was developed with that in mind.

There are several challenges associated with handling large amounts of data: processing power, disk space and RAM memory, alongside hardware constraints and the way in which raw data is presented. The BTS data, as mentioned above, is delivered in the form of individual csv files on a quarterly basis. When dealing with longitudinal studies involving more than one year, it is recommended to opt for a database format instead (e.g. Hadoop, SQL, Spark) to avoid overloading the computer's memory. The R platform allows researchers to access their data in several formats from csv to SQL, showing how this software's interoperability becomes a necessity in a world of dynamic data formats.

SKYNET runs with R or with one of the several available Graphical User Interface (GUI) (e.g. RStudio[49]). To install SKYNET in RStudio, subsequently type in the command line:

```
install.packages(devtools)
library(devtools)
install_github("FilipeamTeixeira/Skynet")
```

More instructions, including how to install it directly from CRAN (https://cran.r-project.org/package=skynet), can be found here: https://github.com/FilipeamTeixeira/skynet.

[48] https://www.r-project.org/about.html

[49] https://www.rstudio.com/products/rstudio/

a) Importing data

SKYNET allows three types of import:

- import_db1b – Imports Comma Separated Files (csv) files downloaded from the BTS website and merges them into an R data frame. It always requires both a Coupon and Ticket file corresponding to the same year and quarter.
- import_t100 – Imports BTS T-100 files in csv format.
- convertRaw – Similar to import_db1b and import_t100, but rather than creating a data frame, it generates a csv file which includes the merged results.

Further details on SKYNET's requirements for data formatting and necessary variables can be found in its ReadMe and Vignette files.

b) Transforming data

When understanding the topological properties of air transport networks and their different nested types is paramount (Barabási & Albert, 1999; Barrat et al., 2003; Guimera et al., 2005; Bagler, 2008; Xu & Harriss, 2008; Rocha, 2017), we are left with the following topologies:

- Route Networks, reflecting real movement of passengers between two points (Coupon concept).
- Metro Networks, reflecting intercity airport agglomerations.
- Origin-Destination, reflecting passenger travel intentions (i.e. the Market concept).

Acknowledging different research questions' needs, we also include extra options such as the possibility to generate both directed and undirected networks, a dichotomization by using the backbone extraction method proposed by Serrano et al. (2009), by filtering per higher weighted edges, etc. When generating a network by running one of the available functions, SKYNET automatically produces an R list with three objects, including a data frame with the OD pairs and other relevant data, an igraph[50] network object and a data frame with all the nodes available on the network and their respective absolute passenger numbers. More information on the functions used, can be found in SKYNET's documentation by typing `help("skynet")` or `vignette("skynet")` in R.

The importance of understanding urban-geographical systems in the context of complex networks has been widely acknowledged, both in general terms (Ducruet & Beauguitte, 2014) as in the case of specific transport networks such as shipping (Ducruet et al., 2010) and air transport (Dai et al., 2018). In this context, researchers are often forced to resort to different programs in order to obtain certain network statistics or in order to simply visualize those same networks. The R ecosystem includes well consolidated packages for both analysis such as igraph (network analysis) and ggplot2[51] (visualization). A simple example with some network characteristics such as small-world (Watts & Strogatz, 1998), scale-free (Barabási & Albert, 1999) and modular community structure (Newman, 2006) was obtained from within SKYNET (Table 1).

[50] http://igraph.org

[51] http://ggplot2.tidyverse.org

Network Characteristics	
Nodes	407
Edges	1235
Small-World	
Cluster Coefficient (Transitivity)	0.09
Average Path Length	2.95
Scale-free	
Kolmogorov-Smirnov statistic	0.0497
Bootstrapped p-value for the K-S statistic	0.2678
Log likelihood for models fitting the degree distribution to log-normal distributions.	-951
Communities	
Modularity	0.293

Table 1 – US air transport network characteristics for Q1 2011

With SKYNET linked to the igraph package, it is possible to easily extract and screen networks based on its characteristics. One of the examples is shown in Figure 5, where by using one of the community detection algorithms found in igraph, it was possible to filter, and then plot the data using SKYNET's plot function.

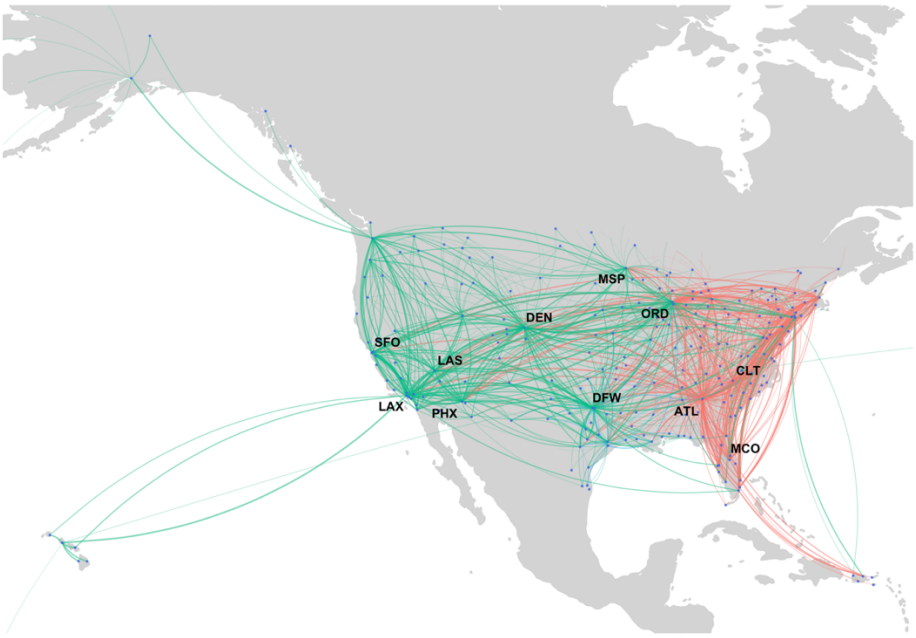


Figure 5 – Airport route network. Year 2011. Nodes shaded by community using the 'leading eigenvector' method (Newman, 2006a).

c) Integration with other tools

As already implicitly shown in the previous section, one of SKYNET's strengths is undoubtedly its capability to integrate with other tools and the potential to be expanded to other sources of data beyond the ones presented in this paper. Due to hardware constraints, researchers are frequently presented with the challenge of handling large files - the BTS data is a case in point. However, being part of the R suite, SKYNET allows a full integration with several database solutions such as

PostgreSQL[52], Hadoop[53] and Spark[54], increasing the flexibility when pre- and post-processing large amounts of data. However, despite SKYNET having the ability of working with longitudinal data without requiring a database solution, it is important to understand that due to R functioning, loading temporal data may a strain on the computer's memory. This makes the discussion between using databases versus importing data a matter of optimum versus convenience.

SKYNET's integration with other tools is not restricted to data management, but also includes options for data analysis and visualization, mostly in the context of network analysis. Despite the clear advantages of using network visualization programs such as nodeXL[55] or Gephi[56], it often becomes cumbersome to transfer data between systems as different formats are required between the different ecosystems. Again, SKYNET being part of the R ecosystem allows its integration with more than 10.000 packages freely available for researchers (CRAN, 2017).

Alongside its ability to communicate with external geographic information systems (GIS) packages (e.g. ArcGIS[57], QGIS[58]) , SKYNET's integration within the R ecosystem, gives access to a large array of (GIS).While they are not able to replace more complex GIS suites, R packages such as sf[59], dodgr[60] and tmap[61], have been consistently and steadily consolidating their position within the GIS world.

The key point here is that the open source nature of the project allows SKYNET to be fully customizable in order to adapt to different research questions, while keeping an often necessary streamlined and standardized data format.

[52] https://www.postgresql.org
[53] http://hadoop.apache.org
[54] https://spark.apache.org
[55] http://www.smrfoundation.org/nodexl/
[56] http://gephi.org
[57] https://www.arcgis.com/features/index.html
[58] https://qgis.org/en/site/
[59] https://github.com/r-spatial/sf
[60] https://github.com/ATFutures/dodgr
[61] https://github.com/mtennekes/tmap

2.4 Applying SKYNET

2.4.1. Hub-and-spoke

Hub-and-spoke forms of organization have been thoroughly studied over the past years (e.g. Campbell and O'Kelly, 2012; O'Kelly, 1998). In order to examine a hub's evolution along with the routes it serves, it becomes necessary to assure data homogeneity in a temporal setting. A common research question within this field of study relates to comparisons between carriers' route structures. This question can be easily answered both visually and numerically, the latter through exploring some more in-depth statistics. Figure 6 shows SKYNET-generated route networks, filtered by carrier, with United Airlines on the left and Southwest on the right. The networks show the edges representing the 10% busiest routes for both carriers, with thicker lines representing higher passenger numbers. In addition to the plot it is also possible to extract more detailed information (Table 2) such as the average fare between airports (itin_fare) and its standard deviation[62] (fare_sd), the yield per mile (itin_yield), and the number of passengers. In this figure we can observe a more compact distribution of routes for United Airlines compared to a sparser distribution for Southwest, which can be attributed to their hub-and-spoke and point-to-point structures, respectively. However, the table generated by SKYNET can help us with obtaining finer details. Table 2 shows that for the routes with the highest passenger volumes, United Airlines has the highest average fares, but the lowest cost per mile tickets (i.e. itin_yield), while Southwest displays the opposite behaviour with lower average fares and a higher cost per mile. A quick glance corroborates the findings by Lium (2009) of hub and spoke networks helping to reduce costs and increasing service frequencies.

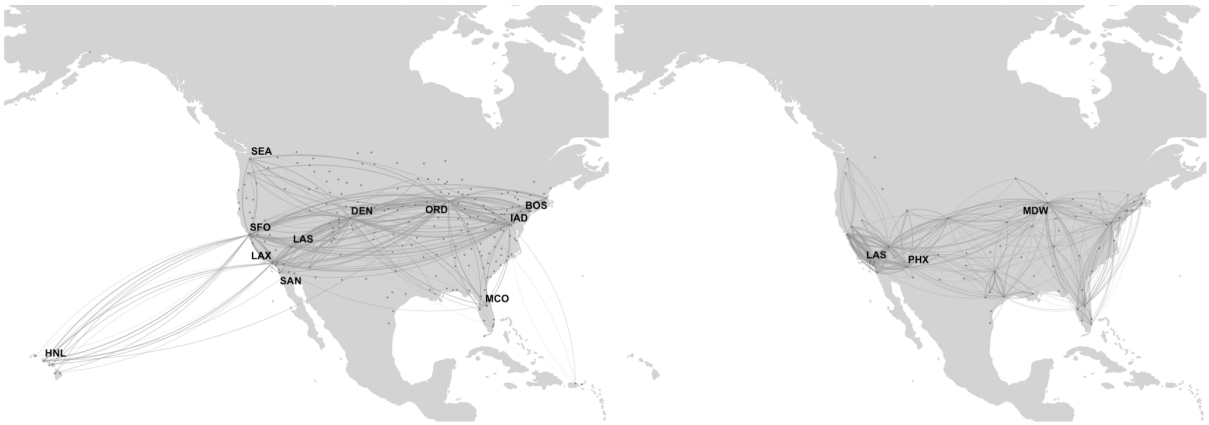


Figure 6 – United Airlines (left) Southwest Airlines (right) 10% busiest routes - Q1 2011

[62] It is important to note that the fare presented in this step is calculated by multiplying the distance between two stops by yield per mile, as the DB1B database only shows the ticket fare, without specifying the cost for the intermediate steps.

Southwest Airlines							
origin	dest	passengers	fare_sd	itin_fare	itin_yield	origin_city	dest_city
HOU	DAL	15439	44.26	77.21	0.32	Houston, TX	Dallas/Fort Worth, TX
DAL	HOU	14920	44.50	76.10	0.32	Dallas/Fort Worth, TX	Houston, TX
LAS	PHX	10869	50.98	60.78	0.24	Las Vegas, NV	Phoenix, AZ
BUR	OAK	10217	49.02	93.71	0.29	Los Angeles, CA	San Francisco, CA
PHX	LAS	10114	51.37	61.28	0.24	Phoenix, AZ	Las Vegas, NV
HOU	DAL	15439	44.26	77.21	0.32	Houston, TX	Dallas/Fort Worth, TX
United Airlines							
origin	dest	passengers	fare_sd	itin_fare	itin_yield	origin_city	dest_city
SFO	ORD	11542	222.16	231.79	0.13	San Francisco, CA	Chicago, IL
DEN	ORD	11190	159.66	162.46	0.18	Denver, CO	Chicago, IL
LAX	SFO	10766	104.98	88.28	0.26	Los Angeles, CA	San Francisco, CA
ORD	DEN	10596	141.48	162.66	0.18	Chicago, IL	Denver, CO
SFO	LAX	10531	112.40	88.89	0.26	Phoenix, AZ	Los Angeles, CA
ORD	SFO	10354	230.39	247.90	0.13	Chicago, IL	San Francisco, CA

Table 2 – Highest passenger volume routes, for Southwest Airlines and United Airlines Q1 2011 – DB1B 10% sample.

2.4.2. 'Pockets of Pain'

In addition to the rather straightforward graphs and tables presented above, SKYNET can complete more comprehensive and complex operations. The flexibility provided by SKYNET becomes evident once the complexity of the research questions increases. In spite of the obvious challenges involved in developing a program that would befit every researchers' needs, the core characteristic of SKYNET lies in having the possibility of answering new questions without writing a large amount of complex code. For example, tackling the post deregulation issues of monopolistic behaviour and the associated asymmetries in air fares across the US urban landscape can easily and directly be drawn from SKYNET.

Table 3 shows different routes, organized by average fare and its variation, denominating them according to the categories presented by Goetz (2002) and Wei (2016): pockets of chaos representing high average fares and high variability; pockets of bliss representing low average fares and low variability, pockets of pain representing high average fares and low variability, and finally pockets of diversity representing low average fares and high variability.

Origin	Dest	Passengers	Fare_sd	Itin_fare	Itin_Yield	AirClusters
LAX	JFK	27674	626.13	557.20	0.211	Pocket of Diversity
JFK	LAX	27151	603.60	539.17	0.206	Pocket of Diversity
SFO	LAX	21842	145.65	157.72	0.422	Pocket of Bliss
LAX	SFO	21614	159.24	157.17	0.426	Pocket of Bliss
LGA	ORD	19684	198.46	246.89	0.308	Pocket of Bliss
SFO	JFK	19559	612.94	541.44	0.198	Pocket of Diversity
JFK	SFO	19506	594.16	532.11	0.197	Pocket of Diversity
FLL	LGA	17389	150.95	209.68	0.172	Pocket of Bliss

Table 3 – 'Pockets of Pain' (Goetz and Vowles, 2000). Q1 2011.

2.4.3. The 'Southwest Effect'

Research tackling the "Southwest Effect" start from the cross-sectional data structure presented in 4.2 and extend It to a setting spanning several years. Vowles (2001) studied the role of Southwest Airlines in altering fares and passenger traffic and the Southwest Effect in multi-airport regions. For these questions, he selected specific routes and looked at both the fares and passenger volume, prior to Southwest entry on that route (1993 Q2), and one year after entry (1994 Q3). As SKYNET was designed to easily import (e.g. by typing import_db1b), and handle data (e.g. make.Path), it is possible to readily print air fares and passenger numbers for a given year. As the DB1B database, which includes air fares, does not include the total number of passengers, we used the T-100 Domestic Market (U.S. Carriers) to complement it. Table 4 shows average ticket fare and number of passengers for a given route, before and after Southwest's entrance on that market.

Origin	Destination	$ prior to entry	$ after year	Pax qtr prior to entry	Pax year after entry
BWI	MDW	$147.25	$73.48	284	55210
BWI	ORD	$188.72	$124.8	90030	121400
SDF	MDW	$220.13	$63.91	16514	39623
CLE	MDW	$147.89	$86.45	47596	94475
MCO	FLL	$117.67	$77.14	27573	49952

Table 4 – Air fares and total number of passengers for Southwest prior and after entry on selected routes. Q2 1993 – Q3 1994.

Pitfield (2008) aimed to understand the impact of the Southwest effect on traffic and market shares. For this analysis, two main corridors (i.e. Washington – Chicago, Philadelphia – Chicago) were chosen. SKYNET allows to quickly produce the required data for such an analysis. Table 5, for example, is the direct output of a simple command, and shows the total number of passengers and market shares for Southwest, United and American Airlines, on the Washington - Chicago (1993-1996,) and Philadelphia - Chicago (2003-2006) corridors, respectively. In the example below, the corridor Washington – Chicago shows a decrease in passenger share for United and American Airlines after Southwest's entry in 1993. The same effect can be seen by the entry of Southwest in 2003 for the Philadelphia – Chicago corridor. One of the main strengths of SKYNET, then, is not just linked to its ability of dealing with several years without requiring an SQL database, reducing the computational complexity and hardware requirements, but with its ability to readily combine data from two databases (i.e. DB1B and T-100) to answer a research question.

Year	Total Pax	Southwest %	United Airlines %	American Airlines %
1993	1076295	2.73	67.77	24.43
1994	3263068	10.77	58.71	20.79
1995	3211786	14.76	54.03	20.90
1996	3194699	15.72	57.48	18.93
Year	**Total Pax**	**Southwest %**	**United Airlines %**	**American Airlines %**
2003	1497428	-	40.36	24.98
2004	1876068	7.15	32.62	24.53
2005	1920027	18.12	32.60	27.18
2006	1873505	21.67	32.27	21.51

Table 5 – Total passengers and total market share for Southwest, United and American Airlines, on the Washington - Chicago (1993 – 1996, top) and Philadelphia - Chicago (2003 – 2006, bottom) corridors.

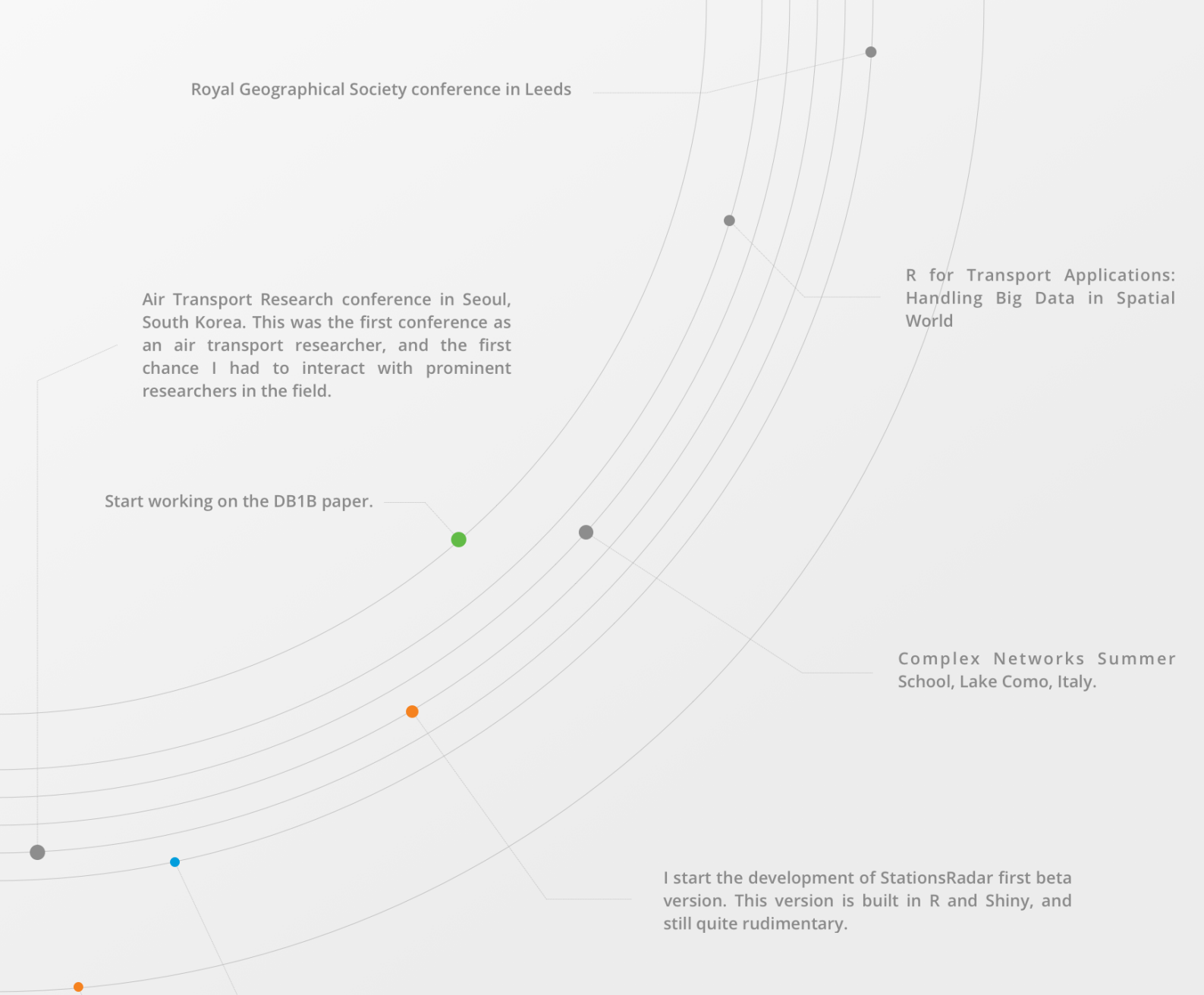
2.5 Conclusion and avenues for future research

In this methodological paper, we have introduced SKYNET, an R package that aims to facilitate air transport-related urban research by showing its flexibility and capacity to overcome some of the challenges faced when analysing some of the urban dimensions of air transport connectivity. For instance, we have shown our software's capabilities to sort, filter and display data, by route, carrier and airport/node, and how the data structure presented by SKYNET could be used to effortlessly reveal different dimensions of unfolding pockets of pain. By generating air transport networks on demand alongside a range of attendant variables (e.g. fare, number of passengers, yield per mile), we automate some of the initial efforts needed to process airline data for urban studies (and other) applications.

However, SKYNET goes well beyond its ability to ease the efforts put into data transformation: its interoperability with, amongst other things, network analysis packages positions it as not just as tool enabling quick access to information in an easy to read-and-use format, but as a starting point for assessing relationships between different sets of data. In 3.2.2 we briefly connect with Neal's (2014b) paper, by demonstrating how within SKYNET, it is possible to connect between network statistics and visualization, which would usually require an external programme.

SKYNET is currently built to handle BTS data and therefore US-centred. Conversely, commercial databases have been preferred by researchers in part due to being presented in a processed and easy to read format, with their free counterparts (e.g. BTS data), being often sidestepped. With this in mind, SKYNET positions itself as a tool within which such questions could be examined as it allows for external data to be imported as long as it follows the same structure (i.e. variables/columns) as in the DB1B/T-100 datasets. As for its ecosystem, despite providing a strong user supported community with data sciences and statistics in mind, R as a programming language is constantly evolving along with most of its packages. Whilst that could possibly be one of its prime weaknesses, it is as well one of its most undeniable strengths.

For the past decade, we have seen a steep rise in the amount of data made available to researchers, often at a small or no cost. However, 'big data' also comes with new challenges for researchers. As data grows both in size and complexity, it is of the utmost importance to find strategies which will allow us to cope with such challenges. However, these challenges are not exclusive to the size of data. As the extensive literature and the datasets available to researchers show, more than ever it becomes important to look at networks as a complex dynamic system with fluctuations ranging from the global network structure to the nodal level, while keeping the inferred conclusions easily open to replication by other researchers. SKYNET does not aim to become the only solution for such problems, but rather an enabling step to tackling complex research questions in the analysis of the urban dimensions of air transport networks.



CHAPTER 3

Revealing Route Bias in Air Transport Data: The Case of the Bureau of Transport Statistics (BTS), Origin-Destination Survey (DB1B)

3.1 Introduction

Air transport research in its various guises analyses the structure of air transport connections, both on their own terms and in their role as a catalyst of wider economic and social developments at various scales (Taaffe, 1956; O'Connor & Fuellhart, 2012; Button & Yuan, 2013; Lin, 2014). The increasing relevance of the research field at large has been fuelled by the observation that both the size and the impact of air transport has been increasing over the past decades (Ishutkina & Hansman, 2008; Air Transport Action Group, 2010). For example, passenger numbers have been soaring, more than tripling from 1.025 billion in 1990 to 3.227 billion passengers in 2017 worldwide (International Civil Aviation Organization, 2017), with IATA forecasting these numbers to double again by 2036 (IATA, 2016). Against this backdrop, researchers have sought to better understand the impact and evolution of air transport, including new and better ways of modelling air transport networks and their many effects.

One of the most critical challenges in air transport research is the uneven availability and formatting of data (e.g. Derudder & Witlox, 2005). While there has been a growth and diversification in data collection strategies over the past few years (Poorthuis & Zook, 2017), air transport researchers still face a limited choice of data sources that may or may not provide information in the desired format and/or detail. When zooming in on researching air transport networks in and of themselves, there are two options. The first option is to build a database from scratch, which in this day and age most often entails collecting web-based air travel data. This can either be done by (1) using or developing a web Application Programming Interface (API) to access a meta-search engine or online route planner and/or by (2) 'screen-scraping' online route planners or meta-search engines (e.g. Grubesic and Zook, 2007). The second option is to use primary datasets and the tools provided to access them. This varies from freely available raw data to sometimes quite expensive databases with bespoke analytical and visualization tools as well as structured APIs (e.g. Google Flights and the Official Airline Guide (OAG) data).There are a small number of exceptions to this bifurcation between freely available versus commercial databases, e.g. the international section of the DB1B database (Bureau of Transport Statistics, 2018), an extra dataset detailing international flights to and from the US. Although in principle freely available, there are some of the restrictions to its usage, such as the requirement of being an US citizen to access the dataset.

In this paper, our focus will be on publicly accessible and arguably some of the most widely used primary air transport datasets: the data provided by the United States Bureau of Transport Statistics (BTS), and its Origin Destination Survey (DB1B) in particular. Analyses based on the DB1B datasets are geographically circumscribed in that – with the exception of the above-mentioned information detailing international flights to and from the US – the data are restricted to information on domestic flights departing from/arriving at United States airports. In addition, the DB1B dataset is a 10% sample of reported tickets rather than a full dataset. In spite of this focus on a sample of US-centred flights, the importance of the DB1B dataset in air transport research cannot be underestimated (Seshadri et al., 2007; Neal, 2010, 2014b; Mao et al., 2015). There are two reasons for this. First, the data 'feeds' parts of some of the other well-known datasets. For example, in the case of the OAG, the DB1B data is used by drawing on the following simple method: "DB1B is a 10% sample of an airline's tickets, then 'adjusted' to estimate 100% of the market by multiplying the data by a factor of 10" (OAG, 2015). Second, the detailed info and the consistent way in which data are gathered make the BTS datasets ideal for air transport research. For example, Neal (2010) uses BTS data to provide an overview of the

use of air traffic networks for urban research by building models of urban-economic flows. Later, he extended this work by differentiating air traffic networks by scale, species and season (Neal, 2014b). Fuellhart (2013) and Brueckner (2014) use BTS data to analyse multi-airport regions (MARs) in the US, disentangling the geographies of offer and demand in these MARs. Brownstein et al. (2006) and Colizza et al. (2007) use BTS data to study epidemiological networks as these are increasingly undergirded by air transport movements.

Irrespective of the diversity of the topics addressed in these and in many other papers, it is clear that one of the drawing cards of the BTS sample as well as other major primary datasets is their alleged 'ground truth'. However, few of the data providers offer detailed info about how their data are sampled and treated. The data quality is therefore sometimes taken to be self-evident. The OAG (2019), for example, self-advertises as "the world's most comprehensive and accurate real-time travel data" without disclosing further details on data collection, treatment and validation. Furthermore, to date there has been little scrutiny of the alleged quality of these datasets (Boyd & Crawford, 2012; Poorthuis & Zook, 2017). The few data quality reports that have been provided tend to be published by the data providers themselves (Zaveri et al., 2012; Strohmeier et al., 2015), with a noteworthy example being a BTS 2005 report on aviation data modernization (US Department of Transportation & Office of the Secretary, 2005). The report determined that 69% of city pairs reported by the DB1B did not meet the Department's accuracy criteria when using enplanement statistics as a validation benchmark. Although this does not necessarily imply that there is a data integrity or structural bias problem with the DB1B data, it does raise questions about how accurate the data are and what the nature of possible biases might be. Because the implications of using inaccurate data may be profound, it is of key importance for air transport researchers to map and understand such possible biases in these datasets.

The purpose of this paper is to explore how potential biases in air transport datasets can be revealed and detailed. Although investigating the sources and impacts of these potential biases would be an equally relevant endeavour, this is a more difficult task. Analysing the sources of biases would require insight into the data collection and processing, but this is complicated because data and their treatment are protected by privacy laws. Meanwhile, assessing the impact of biases on air transport research findings would involve complex issues associated with research replicability and reproducibility. Bearing these limitations in mind, our paper specifically focuses on the (lack of) accuracy of the data rather than on the nature and consequences of these potential biases.

To this end, we develop a methodology that allows identifying possibly biased routes in datasets in an automated manner. Although our focus will be on validating the DB1B data, we present our methodology as a more generic approach that can be used in different contexts and applied to different datasets. To this end, the remainder of this paper is organized as follows. We start by outlining four important factors to consider when working with both the DB1B database and the Air Carrier Statistics (T-100) database (which we use to validate the DB1B database): collection, representation, inter-mutability and nomenclature. We then use descriptive statistics to describe these DB1B and T-100 databases, and propose a Jaccard-like index to identify biased routes. We conclude by demonstrating how route/database biases can impact research by means of a number of straightforward case studies, focusing on our understanding of the position of routes/airports in air transport networks. In a concluding section, we explain how this approach can be adapted to serve as a more generic tool for assessing route bias in air transport datasets.

3.2 The Bureau of Transport Statistics (BTS) datasets

3.2.1. The Origin Destination Survey (DB1B)

The DB1B survey is conducted continuously by all certified US carriers involved in domestic passenger operations. The database covers a "two-tiered" stratified 10% sample, following the 14 CFR part 241 guidelines from the Department of Transport. Data coverage started in 1993, has been updated since, and groups data on a quarterly basis. One of the interesting aspects of this dataset is that it purports to show 'real' passenger movements rather than separate flights. In addition, as the DB1B database displays ticket information, it is possible to look at transfers, fare paid, booking class, intermediate stops and other passenger-related data (Goetz & Vowles, 2000; Vowles, 2001).

The DB1B database comprises three sets of info related with a single entry: Coupon, Ticket and Market (Figure 7). For example, consider the case of a passenger booking a return Ticket from Boston to Atlanta. The first leg of this trip involves a transfer at JFK, the second leg of the trip does not involve a transfer. In this case, BOS-ATL and ATL-BOS would be considered to be the Markets as the Boston-based passenger effectively travels to Atlanta. Meanwhile, the BOS–JFK, JFK–ATL and ATL–BOS segments would be considered to be three individual Coupons. This concept derives from the classic coupon/ticket concept, where tickets were printed rather than available on a digital medium.

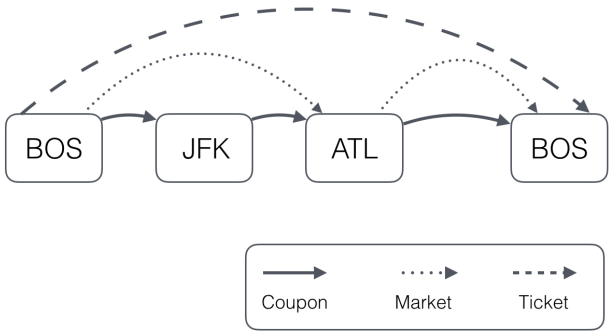


Figure 7 – The concept of coupon, market and ticket in the DB1B database.

According to the Department of Transportation (DOT), the data to be recorded centres on "lifted ticket flight coupons" (i.e. tickets issued by a travel agent, including online ticketing), and includes the following variables: "Point of origin, carrier on each flight-coupon stage, fare-basis code for each flight-coupon stage, points of stopover or connection (interline and intraline), point of destination, number of passengers, and total dollar value of ticket (fare plus tax)" (US Department of Transportation, 2012). As mentioned above, the DB1B dataset is reported as a 10% sample. The randomization is done by carriers by selecting tickets with a serial number ending in zero so that the data collection procedure effectively involves a two-tiered stratified sample.

However, the randomization methods used to build this sample are questioned by the DOT in its report (US Department of Transportation & Office of the Secretary, 2005). For example, it is mentioned that "(s)ince ticket numbers are now assigned by a computer program, the possibility that ticket numbers are assigned for reasons other than randomness arises (…) (A) tour operator might use its block of

ticket numbers to issue all the ticket numbers that end in the same digit to members of a particular tour, resulting in all those tickets being selected for the sample or excluded from the sample depending on which tour was assigned ticket numbers ending in zero". While the deviation from the 10% seemed to be small (i.e. in the 9-11% range) the same report does raise concerns regarding the representation and reliability of information of smaller markets in particular.

Another potential issue associated with the DB1B database can be found in its definition of an Origin Destination (OD) flight. The common definition of an OD flight refers to a passenger traveling from the origin of the trip to the final destination of the trip, including all the intermediate flight stages. However, since the very beginning of the OD Destination Survey, the DOT has used a methodology called Directional Passenger Construction, which uses continuous direction of travel as its definition of 'true' OD. According to this methodology, a passenger is considered to be on a trip as long as the passenger continues in the same direction (i.e. North – South and East – West constitute different direction pairs). For example, on a trip from Albuquerque to Las Vegas with a stopover in Denver, the DOT would break the trip up into two individual trips due to the geographical position of Las Vegas in relation to Albuquerque (US Department of Transportation & Office of the Secretary, 2005). Given the abundance of hub-and-spoke configurations in the organization of many air transport carriers' networks (O'Kelly & Miller, 1994; Campbell & O'Kelly, 2012; O'Kelly, 2016; Park & O'Kelly, 2016), this definition may potentially render more common-sensical notions of a trip, transfer and OD flight inaccurate.

A final challenge associated with this database is associated with the concept of passengers versus passenger trips. As reported by the DOT (US Department of Transportation & Office of the Secretary, 2005), passenger counts are taken to represent passengers scheduled to fly in that quarter. However, the DB1B bundles all travel on a ticketed itinerary in a single quarter (i.e. if a passenger leaves in December and returns in January, the ticket will be reported as if it took place in December and no passenger will be reported for the first quarter of the year). Collectively, these issues may affect how well the sample of observed tickets represent (our conceptions of) the characteristics of actual air travel, even though there are no data regarding how frequent or profound any of these effects are.

3.2.2. The Air Carrier statistics dataset (T-100)

In principle, directly validating sampled data relies on having a full dataset. This full dataset is not available, which implies that we have to resort to another dataset: the BTS Air Carrier statistics dataset (T-100). The T-100 dataset contains domestic and US-related international airline market and segment data. Certificated US air carriers (i.e. carriers with granted approval by the National Aviation Authority allowing the use of their aircrafts for commercial purposes) (Bureau of Transport Statistics, 2019b), have to report air carrier traffic information on a monthly basis, using the so-called Form T-100 to the Office of Airline Information, Bureau of Transportation Statistics (Bureau of Transport Statistics, 2019a).

The T-100 differs from the DB1B in several respects, starting from its time structure. In contrast to the DB1B, the T-100 database is grouped per month, while it represents a full sample and reports flights rather than individual or group tickets. Although it is less fine-grained than the DB1B dataset on a number of fronts, T-100 data is also often used in air transport research. By including all boarded passengers, it can help answering questions related with airport and carrier competition (Button & Lall, 1999; Dobruszkes & Van Hamme, 2011; O'Kelly, 2016; Song & Yeo, 2017), and it has also been used

for better understanding route level variations within multi-airport regions (Fuellhart, 2007; Fuellhart et al., 2013).

The differences between the DB1B and T-100 datasets imply that using the latter full dataset to validate the former sampled dataset is not a straightforward exercise, a situation that is further complicated by the subtle differences in the terminology used to identify a Segment and a Market pair. The T-100 database identifies a Segment as a non-stop flight, including diversions, flag stops, tech-stops, emergency landings, etc. This data is not flight number driven and referred to as "transported data". However, Market data are flight number driven, which implies that if the flight number changes the market stops. Market data are often in research referred to as "enplanement data". Figure 8, using fictional flight numbers, shows how a flight can be both part of the Market data and the Segment data. Flight, UA01 (BOS – ATL with a stop in JFK) and UA02 (ATL – BOS), both represent a market, as the flight number stays the same. Following a similar example posted online on the BTS website (Bureau of Transport Statistics, 2019c), we consider 250 passengers to take a flight from BOS, of which 200 deplane in JFK while 70 new passengers board and continue to ATL where the 50 remaining and the 70 newly enplaned passengers deplane. The T-100 Market/Segment dataset would respectively show:

BOS – JFK: 200/250 passengers
JFK – ATL: 70/120
BOS – ATL: 50/(not available)

A key aim of this paper is to explore the actual existence of routes in the DB1B dataset. While it is clear these cannot be directly found in other data sources, it is possible to indirectly derive these by combining the T-100 Market and the T-100 Segment data (Figure 9) which jointly represent all possible route combinations.

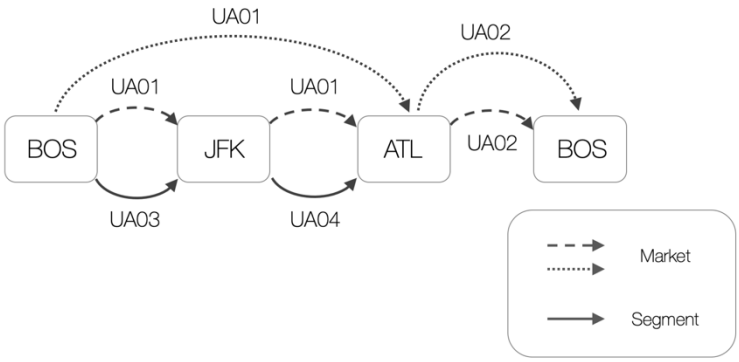


Figure 8 – T-100 concept of Market and Segment (flight numbers are fictional).

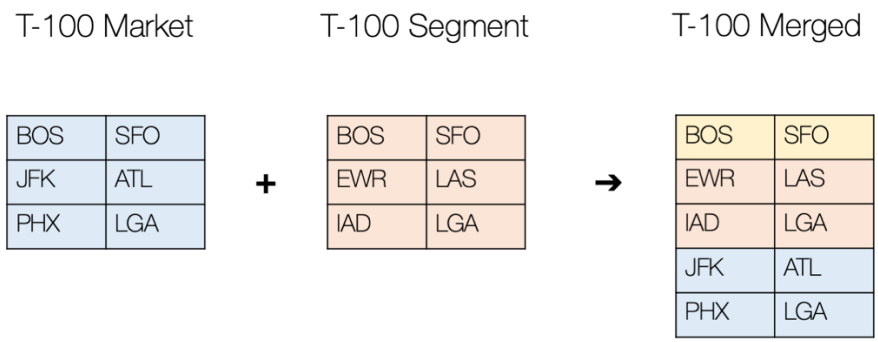


Figure 9 – Schematic showing method used to combine T-100 Market and T-100 Segment datasets.

Taken together, then, the data collection is broadly similar to the one used for DB1B, albeit that there are clearly also some major differences regarding the collection and segmentation of the data. For example, the T-100 database, rather than being a 10% sample, includes all available data. Other major differences relate to what is reflected in the data: whereas the DB1B reflects tickets that were effectively bought (i.e. the data is collected before a flight occurred), the T-100 reflects the exact number of passengers who actually boarded a flight (i.e. the data is collected after a flight occurred). However, as individual information is not reported in the T-100, it is difficult to infer information regarding transfers or ticket fares.

3.2.3. Comparing the DB1B and T-100 datasets

Partly due to the differences in data collection procedure, the structure of the data, and the terminology, it is difficult to directly compare the DB1B and T-100 databases. Or, when cast in the context of this paper: the less detailed but full T-100 data cannot be directly used to validate the more detailed but sampled DB1B data. Fortunately, the key challenge in establishing connections between the two datasets is restricted to comparing passenger volumes on the same route. Although the terminology and data structure does not simply translate between these two datasets, it is possible to develop a method measuring the presence of a route throughout time in the DB1B dataset and comparing it with its presence in the T-100 dataset. However, this requires establishing a terminological comparison of the Segment/Market concepts in the datasets. It is important to keep in mind that we will interchangeably use 'route' as a way of referring to all types of passenger movements (i.e. both take off–landing and Origin–Destination pairs).

Figure 10 shows a fictional passenger traveling between BOS and LAX. This passenger took two flights (i.e. UA01 and UA06), and transferred in ATL. Meanwhile, the first flight (i.e. UA01) had a stop in JFK, albeit that the passenger did not disembark there. In the DB1B dataset, this passenger – if part of the sample – would be represented by two segments (i.e. BOS-ATL and ATL-LAX) and one market (i.e. BOS-LAX). However, in the T-100 dataset, the same passenger would be represented by three segments (i.e. BOS-JFK, JFK-ATL and ATL-LAX), and two markets (i.e. BOS-ATL and ATL-LAX). In light of this, it may seem tempting but nonetheless incorrect to infer that the DB1B Coupon and the T-100 Market are simply interchangeable, as the T-100 refers to 'enplaned passengers' whereas the DB1B refers to 'revenue passengers' (Office of the Secretary - Department of Transportation, 2019). However, what we can be directly compared between both datasets are the routes represented. At the same time, it is important to understand that such a comparison is only possible when both the T-100 Market and Segment data are combined as per Figure 6. This comparison relies on the principle

that the T-100 Segment shows point-to-point routes, whereas the T-100 Market shows market segments. By merging the two datasets, we capture all available routing options, and can compare these with information derived from the DB1B sample.

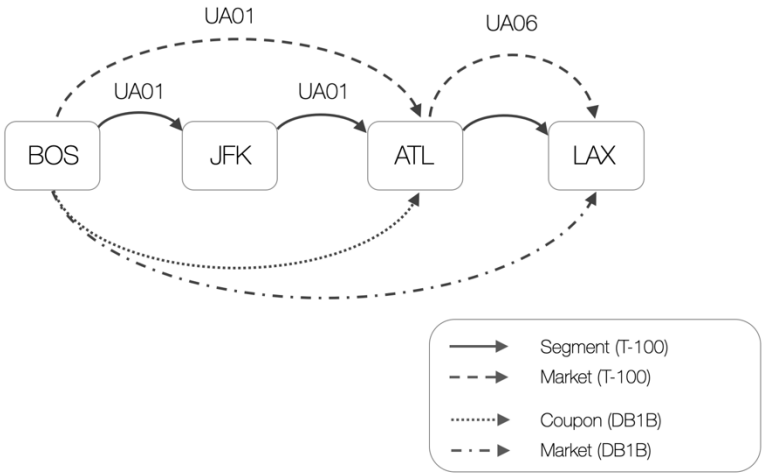


Figure 10 – Segment and market concept in the DB1B and T-100 datasets

3.3 Testing the randomness of the DB1B 10% sample

To compare both datasets, we work with 11-year samples between 2005 and 2015 on a quarterly basis. The data are processed into networks using the R package SKYNET (Teixeira & Derudder, 2018). Both networks were transformed so that they are comparable in terms of their nodes and edges: we merged data as to produce matrices with undirected edges featuring the exact same sets of origins and destinations, carriers, and timeframes. Because the T-100 generated network includes the full sample of airports and routes, we intersected both networks to generate a T-100 adapted sample by retaining the largest connected component of the network. Due to the 10% sample characteristics of the DB1B, we multiplied the number of passengers by 10 to have a number that is comparable with the T-100. As routes represented in the DB1B database will in theory amount to 10% of the T-100 values, T-100 routes with less than 10 passengers per quarter are not likely enough to appear in the DB1B to make a meaningful comparison, and these were therefore excluded from the analysis.

3.3.1. Descriptive Statistics

Figure 11 shows the number of edges (routes and passenger volumes) and nodes (airports) over time. The picture emerging here suggests a broadly consistent pattern for both databases, with the number of airports slightly decreasing and the number of passengers slightly increasing over time, except for the number of routes which remain stable in the T-100 database and decrease in the DB1B database. As can be expected, these longer-term trends are interspersed with cyclical year-long patterns with the first and fourth quarters having lower number of passengers (Bureau of Transport Statistics, 2017). Nonetheless, the number of routes is considerably higher and the number of passengers considerably lower in the DB1B dataset than in the T-100.

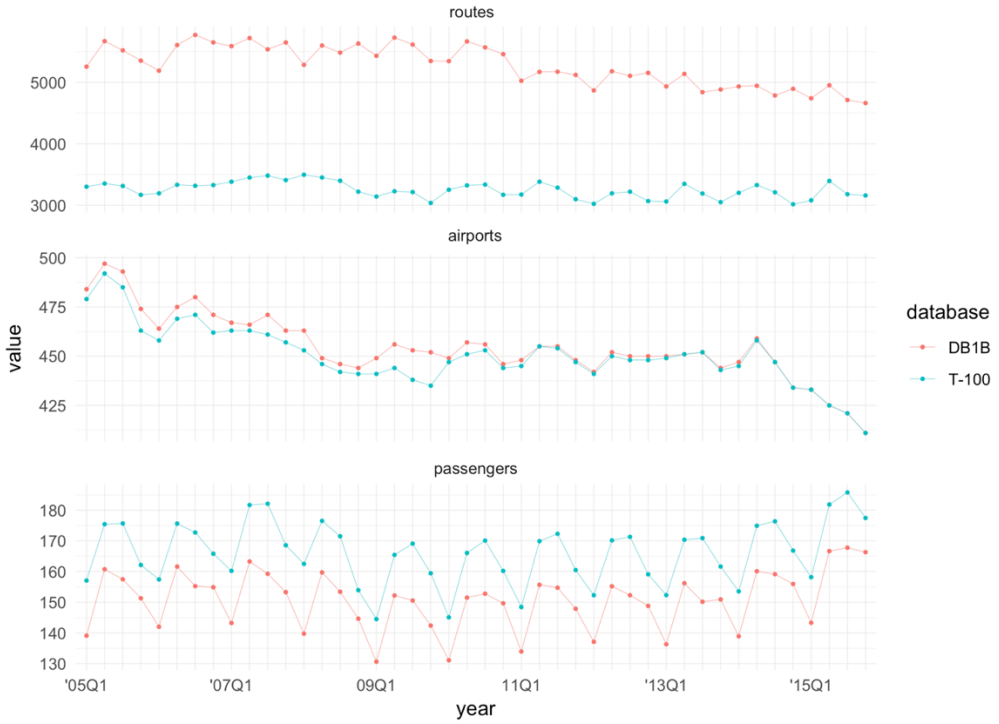


Figure 11 – Quarterly number of routes, airports and passengers (thousands) between 2005-Q1 and 2015-Q4 in the DB1B- and T-100 datasets. Note that for the airports graph, data between 2014 Q1 and 2015 Q4 is similar for the DB1B and T-100 databases.

After testing for the presence of normal distributions by running a Shapiro-Wilks test, we calculated Pearson's correlation coefficients between the same variable in both databases (e.g. number of airports per quarter in the DB1B against the same value for the same period in the T-100). We observe a strong correlation at the level of the airports ($r = 0.9695$, $p < 0.001$, Figure 12b) and passengers ($r = 0.950$, $p < 0.001$, Figure 12a). However, this does not hold at the route level ($r = 0.547$, $p < 0.01$, Figure 12c).

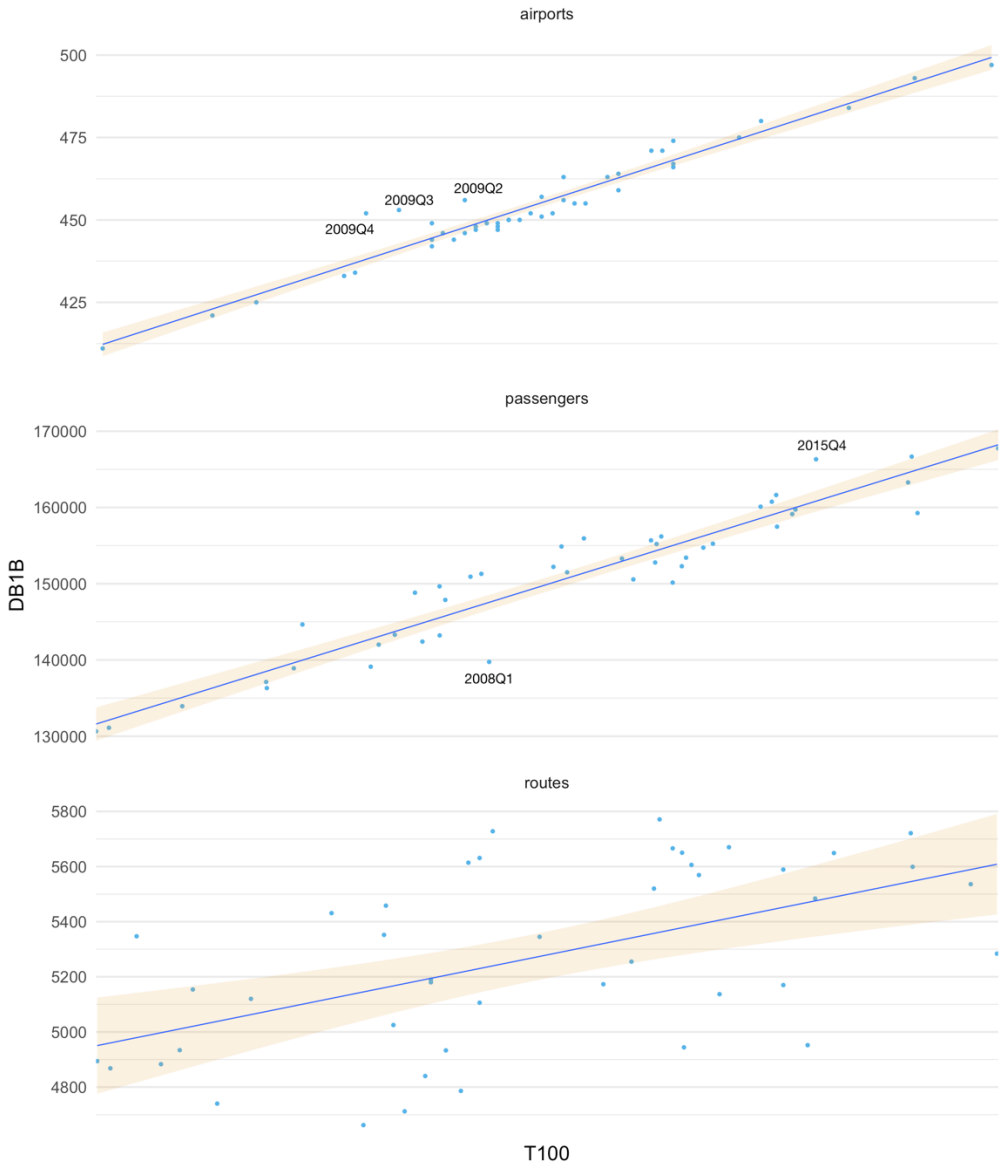


Figure 12 – Correlation analysis of airports, passengers (thousands) and routes per quarter for T-100 and DB1B (CI = 95%).

Taken together, it is clear that while the number of passengers and airports are roughly consistent in the DB1B and T-100 datasets, this is not the case for the number of routes. In addition to the low correlation, the number of routes in the DB1B dataset is consistently higher, which is implausible for a random sample (i.e. the number of routes remains consistent in the T-100 while it decreases with about 10% in the DB1B during the period under analysis). Further analysis shows that when intersecting the two databases at the route level, we see that an average of 38% (min 33%-max 43%) of the routes found in the T-100 database cannot be found in the DB1B database in the corresponding period (Figure 13). This pattern becomes even more compelling when selecting only routes with a number of passengers higher than the 75% quintile (which can be hypothesized to be more consistent and/or

less prone to random effects) because even in that case an average of 15% (min 9% - max 22%) of the DB1B routes are missing from the T-100 database (Figure 14).

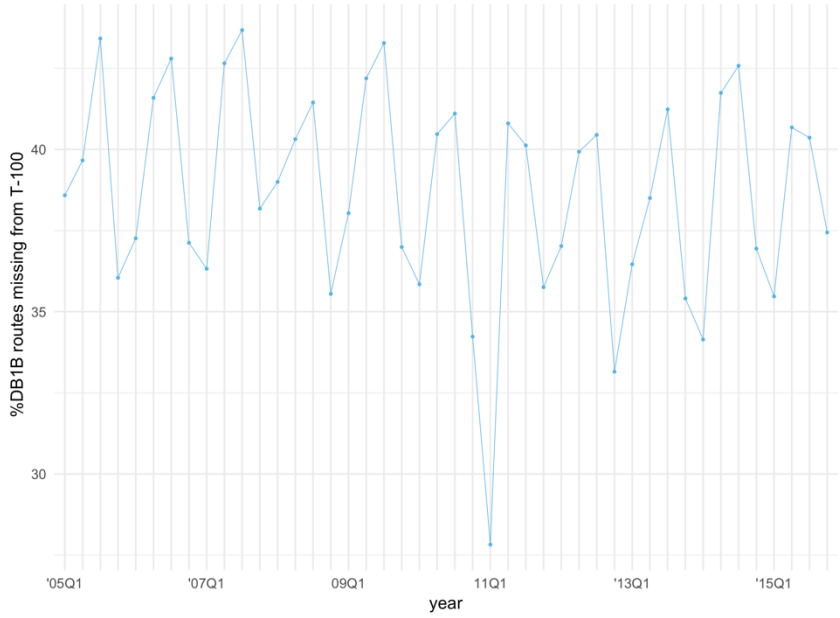


Figure 13 – Quarterly percentage of DB1B routes not present in the intersection between the DB1B and the T-100 Segment.

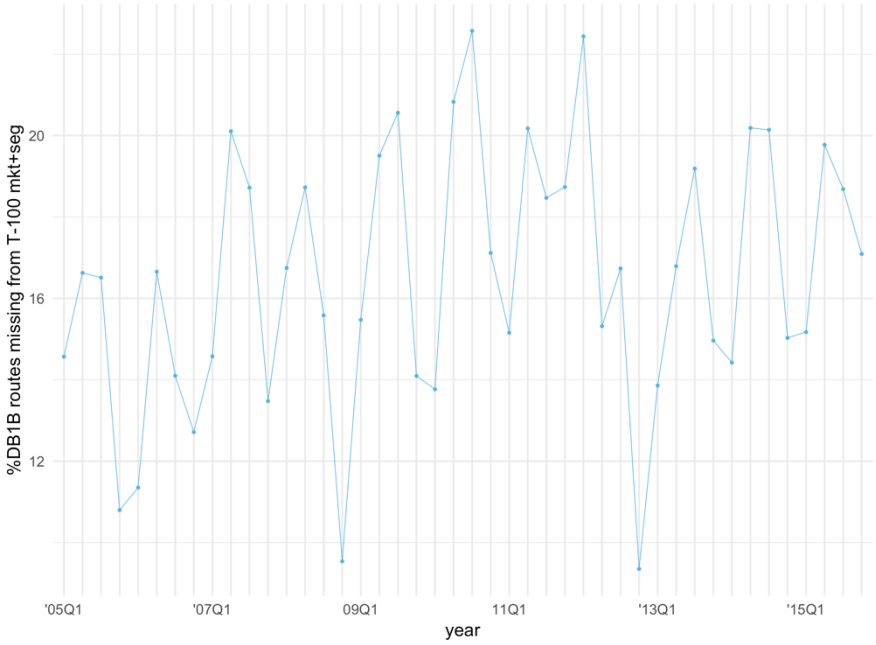


Figure 14 – Quarterly percentage of DB1B routes no present in the intersection between DB1B and T-100 (combined Market and Segment) for the passenger quintile above 75%.

3.4 Conceptualising a Jaccard-Like index – The Route Equality Ratio

The above exploratory assessment of the data revealed that the main differences between the T-100 and DB1B databases are route related. As the combination of the T-100 Segment with the T-100 Market database should in theory display all available routes, it can be inferred that any route exclusive to the DB1B should be considered to be inaccurate. With that in mind, we can assess and compare the presence of routes in both datasets over time in more detail. To this end, we develop a Jaccard-like index to identify how (un)evenly routes are covered in both datasets. In other words, our aim is here to capture and subsequently compare the presence of a route in the DB1B database with its presence during the same period in the T-100 database. To this end, we develop the Equality Ratio (EQR) associating route frequency (RF) between any pair of airports in both datasets.

The calculation of the frequency of route presence RF consists of two complementary parts. The first parts captures yearly frequency per quarter (i.e. to capture route seasonality), while the second part captures the quarterly presence per year (i.e. to capture dispersion over years). In Equation 1, we define Ny_{ij} as equal to 1 and Nq_{ij} equal to 0.25 when a route is present for any given pair ij of origin destination airports in both conceptions, respectively. As most routes tend to have a seasonal dimension (Burghouwt & de Wit, 2005; Mao et al., 2015; Sun et al., 2015; Rocha, 2017), a heavier weight is given to the yearly presence per quarter (i.e. Ny_{ij}). The frequency of route presence between airports i and j (RF_{ij}) can then be calculated as follows:

$$RF_{ij} = \frac{\sum_{q \in Q}(\sum_{y \in Y} Ny_{ij})}{|Y|} + \frac{\sum_{y \in Y}(\sum_{q \in Q} Nq_{ij})}{|Y|}$$

With:

- RF_{ij} = Route frequency for routes between airports i and j
- Y = Range of years
- Q = Range of Quarters
- $\boldsymbol{Ny_{ij}} = \begin{cases} 1, & if\ the\ route\ exists \\ 0, & if\ the\ route\ does\ not\ exist \end{cases}$
- $\boldsymbol{Nq_{ij}} = \begin{cases} 0.25, & if\ the\ route\ exists \\ 0, & if\ the\ route\ does\ not\ exist \end{cases}$

Equation 1 – Route frequency

RF_{ij} ranges from 0 (never present in a dataset) to 5 (consistently present in a dataset), and is used to calculate the Equality Ratio, which is given by:

$$\mathrm{EQR_{ij}} = \begin{cases} \dfrac{RF(\mathbf{DB1B})_{ij}}{RF(\mathbf{T-100})_{ij}} & for\ RF(\mathbf{DB1B})_{ij} < RF(\mathbf{T-100})_{ij} \\ -\dfrac{RF(\mathbf{T-100})_{ij}}{RF(\mathbf{DB1B})_{ij}} & for\ RF(\mathbf{DB1B})_{ij} > RF(\mathbf{T-100})_{ij} \end{cases}$$

With:

- EQR_{ij} = Equality Ratio for routes between airports i and j.

EQR_{ij} ranges from -1 to +1, with a value of 1 representing a perfect match (i.e. consistent route presence/absence in both datasets) and negative values indicating a higher frequency of route presence in the DB1B than in the T-100. If the DB1B dataset would be an unbiased sample, we would expect high positive values converging on a value of +1.

Table 6 shows some examples of routes and their associated RF and EQR values. If we take the example of ABE-CLE, which has an RF equal to 2.39 in the DB1B database, we can infer that the route tends to be fairly but not consistently present. In the T-100 database, this route has a RF equal to 3.5, which implies that the route has been almost always present across years and seasons. For the ABE-CLE example, this produces an EQR equal to 0.68, which captures that the frequency for that route is not the same for the two databases. Some of the other routes (e.g. ABE-ATL) are consistently present in both datasets and have the expected value of EQR of 1, but it can be seen that for quite a large number of routes the EQR is indeed not equal to 1. A systematic appraisal of EQR across airports allows assessing how well/poorly both databases match.

Origin	Destination	DB1B frequency	T-100 frequency	Equality ratio
ABE	ATL	5	5	1
ABE	AVP	2.72	1.70	-0.62
ABE	BOS	0.1	1.1	0.1
ABE	CLE	2.39	3.5	0.68
ABE	CLT	5	5	1
ABE	CVG	1.7	1.7	1

Table 6 – Route frequency in DB1B and T-100 and the associated EQR, for first 6 alphabetically ordered routes by origin and destination.

In Figure 15, which shows the distribution of EQR values for all sampled routes between 2005 and 2015, we can observe that over 40% of the routes (shown in blue), have an EQR lower than 0.85. Even though an EQR of 1 is by far the single largest value, which suggests that there is indeed a fair share of routes that is consistently covered in both datasets covering 45% of all routes, there is no convergence on 1. Furthermore, there is a substantial number of routes with an EQR < 0. Overall, then, we find that routes are either consistently covered (the grey bar to the right in the EQR range) or exhibit different and seemingly almost random presences in both databases (the distribution in the remainder of the EQR range). As can be expected, this is slightly less the case for the busiest routes: routes with RF values of 5 – think: JFK-LAX or ORD-SFO – do more often result in EQR values of 1.

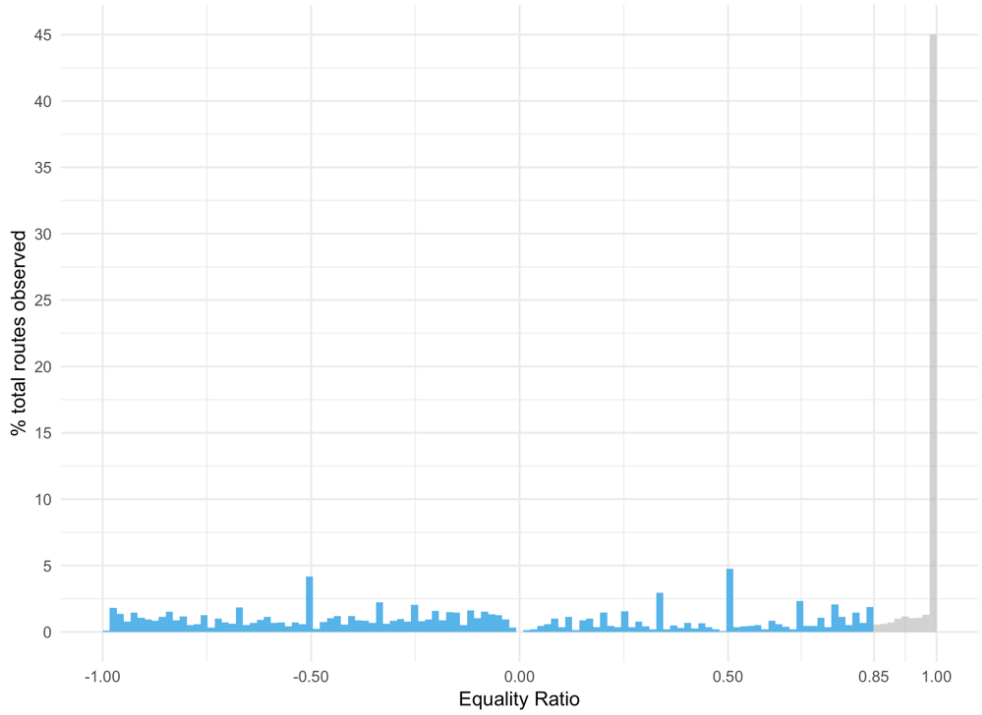


Figure 15 – Histogram of percentage of routes observed by EQR for all sampled routes between 2005 and 2015.

3.5 Assessing the impact of using biased data

Although almost half of the routes in the DB1B sample are indeed coherently present in time and space, this also implies that more than half of the routes in the DB1B sample do not coherently reflect actual routing in space and time. In addition, there is no convergence in EQR on 1, which suggests that some routes in the DB1B dataset apparently have no reliable real-world referent. To describe the potential impact of this issue, we explore what it this may mean in practice for two types of air transport research: airport focused research and route focused research.

3.5.1. The potential impact on airport focused research

While it can be expected that some of the lower passenger volume routes may not be present in the DB1B due to its 10% sample characteristics, below we will demonstrate that even some of the busiest routes are not always represented in both DB1B and T-100 databases. Table 7 shows the mean EQR for seven airports of varying importance alongside other information that is relevant to reflect on the potential impact of the 'ghost routes', i.e. routes present in the DB1B but not in the T-100 database. The table shows, for example, that an airport such as LAX with more than 25 million passenger departures per year in 2015 had almost 50% of its unique routes in the DB1B dataset not regularly matching the routes present in the T-100 database.

Airport	Passenger (departures for 2015 in thousands)	Equality ratio (mean)	Equality ratio (standard deviation)	% routes with EQR = 1	Total number of unique routes (incoming and outgoing)
Hartsfield-Jackson Atlanta International (ATL)	44319	0.801	0.510	72.48	447
Chicago O'Hare International (ORD)	30661	0.806	0.491	74.29	389
Dallas/Fort Worth International (DFW)	27914	0.761	0.563	71.39	395
Los Angeles International (LAX)	26854	0.514	0.699	48.61	323
Denver International (DEN)	25521	0.684	0.628	64.07	437
Phoenix Sky Harbour International (PHX)	20773	0.623	0.611	55.05	287
Charlotte Douglas International (CLT)	20719	0.805	0.509	74.23	291

Table 7 – EQR for 7 busiest airports by volume of departed passengers (domestic flights only) in the US (2005-2015)

There are several papers in the literature dealing with the impact and relevance of specific routes for airports (Goetz, 1993; Goetz & Vowles, 2000; Wei & Grubesic, 2016). Analysing an airport that has a substantial number of 'ghost routes' entails that the research question is tackled with skewed data and may subsequently affect the researcher's final conclusions. Research on Multi Airport Regions (MARs) is one research area where airport-level variations in routes are tackled (Fuellhart, 2007; Fuellhart et al., 2013, 2016; Brueckner et al., 2014). Brueckner (2014), for example, provides a methodology looking at the effects of competition on airports in MARs by drawing on DB1B data. Although it is not

clear which routes were used and how these are affected by the pattern described here, and although we acknowledge that Brueckner (2014) filtering the data by removing routes shorter than 200 miles and with less than 10 passengers per each way may have tackled some the issues raised here, some of the airports analysed do display a high percentage of routes with a low-quality EQR. Meanwhile, Fuellhart (2013) looks at route-level passenger variations within three multi-airport regions (i.e. Boston, San Francisco and Washington). While the data he used was retrieved from the T-100 database, it is nevertheless clear that there would be an impact if DB1B data would be used to answer similar questions. For example, we can observe that all the largest airports within these three areas have a high percentage of low-quality routes (i.e. BOS – 18,7%, SFO – 32%, IAD – 23,04%).

3.5.2. The potential impact on route focused research

Our analysis reveals that between 2005 and 2015 there are 782 unique routes with more than 137 passengers per day (one way) where their frequency in the DB1B database is higher than in the T-100 (the threshold was calculated by selecting routes with passenger volumes above the second quintile).

At the level of the routes, there are two important elements to consider. The first one resides in the assumption that as the DB1B is a 10% sample, it is expected that some – especially: smaller – routes may end up being missing from that database. However, we can assume that the opposite is in principle not possible as the T-100 Segment and Market should include all available non-stop or flight number dependent routes. The second element is the number of passengers shown for a particular route. While some small differences are expected due to the characteristics of the sampling, it is hard to assess the number of passengers represented on a route due to the way in which the data is collected (i.e. ticket collected vs passengers flown), but most importantly due to not being able to compare the number of passengers in the 10% sample with a full dataset. The literature shows several papers studying the impact and relevance of routes on the economy, in terms of competition, or as a way of understanding people's movements (Dresner et al., 1996a; Vowles, 2001; Neal, 2014b). In such research routes are not aggregated per airport, but scrutinized individually which makes research even more vulnerable to the effects of routes being overrepresented in the DB1B database. With this in mind, it becomes crucial to understand if a route is over-represented (i.e. its frequency is higher in the DB1B than in the T-100 database). It is nevertheless important to reiterate that we do not aim to directly compare the number of passengers between the DB1B and the T-100 databases, but rather to show the existence of potential biases, i.e. (1) where no results are found in the T-100 datasets but results are found in the DB1B, and (2) where the number of passengers found in the DB1B is higher than the number seen in the T-100 for the same route and period.

To explore this issue, we selected three routes (i.e. BPT-IAH, OTH- SFO, MSP-PSP) based on our preliminary finding that their presence was higher in the DB1B than in the T-100 database. When looking at the BPT (Beaumont/Port Arthur, TX) – IAH (Houston, TX) route, we can see that for 2012 Q4, 2013 Q1, 2014 Q1, 2014 Q3, 2015 Q2 and 2015 Q4, there were no results in the T-100 database, while the DB1B does show results for that route. While the passenger average per month is low (i.e. approximately 980) it represents nevertheless a good case study due to the lack of matching results for the same period and route in the T-100 database. in 2012 Q4, the route showed 640 passengers in the DB1B. In the case of the route OTH (North Bend/Coos Bay, OR) – SFO (San Francisco, CA) we can see a similar pattern. Despite having a more consistent presence, it is possible to observe in both databases that this route was being served by SkyWest Airlines (OO), but it is only in the DB1B that this same route emerged as being operated by United Airlines (UA) and Continental Airlines (CO). Although

it is known that SkyWest often flies for United (Wickham, 2011), this does not explain United's absence in the T-100 database. However, it is important to notice as well that 2011 Q4 shows 7297 passengers flying the OTH – SFO segment for the T-100 and 9090 passengers for the DB1B. Assuming that these are direct flights it is difficult to understand the extra carriers and the different number of passengers shown.

The issue becomes even more compelling but complex, when analysing our last example: the route between MSP (Minneapolis/St. Paul, MN) and PSP (Palm Springs, CA). Table 8 shows a rough comparison between the three datasets. The presence of Northwest Airlines (NW) in the DB1B database and its absence in the T-100 can easily be explained by its end of operations on the 31st of January 2010 and its subsequent merger with Delta Airlines (DL) (Luo, 2014). However, Sun Country Airlines (SY), with its main hub in MSP, displays more passengers for a segment which is similar in number of passengers for both the T-100 Segment and Market datasets. On Sun Country Airlines website (Sun Country Airlines, 2018), it can be read that they offer non-stop flights between both airports except for the earlier-mentioned "roll-over" behaviour where a passenger can see its ticket being changed to another carrier.

Operating Carriers	AA	DL	NW	SY	UA
DB1B	140	15850	430	9930	-
T-100 Segment	-	17063	-	9633	-
T-100 Market	163	16677	-	9633	291

Table 8 - passengers for route MSP - PSP, 2010 Q1

This uneven presence of routes can, for example, have an impact in research that looks at route competition in the US (Morrison & Winston, 1990; Bania et al., 1998; Borenstein & Rose, 2002). While airline mergers and "roll-over" effects can impact the observed data (e.g. the already mentioned NW merger with DL), some of the overrepresented values in the DB1B database cannot be directly and easily explained. Nonetheless, tagging routes based on their EQR can be used as a means of understanding if they can be directly used in research or if further scrutiny is needed.

3.6 Discussion and final remarks

The primary focus of this paper has been to explore the presence of potential biases in the DB1B database, the source of some of the most commonly used datasets in air transport research. To this end, we developed a methodology that allows identifying possibly biased routes: a Jaccard-like index was proposed to compare route presence in the DB1B data against a route presence database derived from T-100 data. Importantly, this approach implies that our methodology has broader purchase in that it can be cast as a more generic approach to validate route presence in different contexts and applied to different datasets: our proposed EQR algorithm can – in this or an amended form – be used to identify potentially biased routes in other databases as well. To the best of our knowledge, the DB1B database has not yet been validated with the partial exception of an earlier and equally critical BTS self-assessment. One reason for this is that validating the DB1B database is not straightforward because of a range of differences with other BTS datasets. As we have discussed in this paper, there are four important factors to consider when comparing the BTS databases: Collection – how are the tickets collected and sampled? Representation – does the data reflect the final results (i.e. does it represent a passenger boarding a flight or the ticket bought by a passenger)? Inter-mutability – can data be transferred between the DB1B and T-100 datasets? And nomenclature – do variables with the same name mean the same thing in both datasets? By coherently considering the above factors, it became possible to validate the DB1B database using T-100 data.

Our main findings are that (1) although roughly half of the routes in the DB1B dataset are consistently present in the T-100 dataset there are also many inconsistent routes, and (2) that although this issue is present across routes and airports this is somewhat less pronounced for important routes and airports. Our research is able to identify some of the issues with the DB1B data, but is if course not able to shed light on the source(s) of the issues or the actual impact on previous or future research, even though the BTS self-assessment and some of the route level bias examples discussed in the previous section offer some suggestions. That said, the purpose of this paper has not been to invalidate the DB1B database or its potential relevance: its level of detail, consistency, wide longitudinal range, and free availability alone imply that it remains a premier data source for air transport researchers. Other data sources often costs tens of thousands of dollars or are locked behind confidentiality agreements for data usage (Huang et al., 2013), making it almost impossible to unlock these for research purposes. As the impact on previous research remains unclear, follow-up research offering an in-depth exploration would complement this paper. Here, we deliberately focused on revealing the presence of biases, and in the process equipping air transport researchers with the tools to assess bias in databases. We hope this serves as a constructive starting point for further discussions on data quality and data availability in air transport research. Follow-up research can thus also focus on potential solutions to the issues raised here. One preliminary suggestion is to use our method to isolate potentially biased routes and the validate these using secondary sources. For example, researchers could flag routes with an EQR < 0.85, and then use airlines' websites or other secondary sources to cross-check the actual presence of the route. Another possibility is to add our method to existing frameworks (e.g. TensorFlow, Keras, Torch), mostly in the domain of Machine Learning, to build a fuller dataset (e.g. Abadi et al., 2016; Chollet François, 2015; Collobert et al., 2016). As our method looks at route presence and potentially frequency rather than passenger volume, EQR scores can be used to label routes in order to be later used by machine learning algorithms to "fill in" missing values by using different imputation methods

DB1B paper is published in the journal of air transport management

Statistical Analysis for Space Time data in R, Lisbon, Portugal.

I start working on StationsRadar v2.0. This version is written in JavaScript and Vue.js, and supported by R and Shiny.

Final version of StationsRadar is released to the public, in time for Freke Caset PhD defense.

I start working on the MARs paper.

European Transport Conference in Dublin, Ireland.

CHAPTER 4

Spatio-temporal dynamics in airport catchment areas: The case of the New York Multi Airport Region

4.1 Introduction

The analysis of large metropolitan regions has become a major field of research across scientific disciplines (Harrison & Hoyler, 2014; Yeh & Chen, 2020). One of the transport geography research agendas within this literature focuses on the region-wide provision of air transport connectivity, which given the (market) size of these metropolitan regions often occurs through multiple airports. In such Multi Airport Regions (MARs), passengers have – to some degree – a choice between airports (Derudder et al., 2010). For example, in an archetypical MAR such as the San Francisco Bay Area, passengers *inter alia* need to consider the accessibility of airports (e.g. Oakland Airport (OAK) being relatively more accessible from the region's east); the nature of connectivity supply at airports (e.g. San Francisco International Airport (SFO) offering more direct international connections); and a broad range of miscellaneous factors potentially determining the utility of airports, airlines and airport-airline combinations, such as quality of service, pricing, loyalty programs, on-time performance, and onward connectivity (Pels et al., 2001; Hess & Polak, 2006; Thelle & Sonne, 2018). Furthermore, planning and policy frameworks may also have an impact on airport choice in a MAR, as for example shown by the late-evening curfew on take offs and landings at Santa Ana impacting flight availability across the Los Angeles metropolitan area (Fuellhart et al., 2013).

Although the MAR concept and research into some of its key dimensions have been around for a while (e.g. Harvey, 1987; de Neufville, 1995), this field of research is becoming increasingly germane (Bonnefoy, de Neufville & Hansman, 2010). There are a number of urban and regional processes that produce new geographical settings in which MARs develop, such as the fast growth of multiscalar urban clusters in China (Liu et al., 2018; Yang et al., 2016). Meanwhile, the observation that in some parts of the world borders have become less of an obstacle for passengers has led to the notion of cross-border MARs, as shown in Paliska et al.'s (2016) analysis of the Italo-Slovenian Upper Adriatic region. The MAR concept is also becoming institutionalized, especially in the United States: regional codes are used in booking systems and sometimes explicitly recognized by IATA, as for example shown by the single codes for the San Francisco Bay Area (QSF) and New York (NYC) airports. And finally, even in regions that may not be on the map of 'classical' MARs, MAR-like research questions emerge. Fuellhart, (2007), for example, showed airport substitution patterns for Harrisburg International Airport (MDT), located in south-central Pennsylvania, towards various more-or-less proximate airports.

From a transport-geographical perspective, a thorough understanding of airport choice in MARs is rooted in an understanding of airports' *catchment areas* (e.g., Fuellhart, 2007; Lieshout, 2012; Paliska et al., 2016). In this paper, we aim to contribute to this literature by exploring a dimension of MAR airport choice and therefore MAR airports' catchment areas that has previously remained under the radar: their spatio-temporal dynamics. Irrespective of the empirical or analytical focus, existing MAR research tends to conceive airports' attractiveness and catchment areas as spatio-temporally static. This can to a large degree be traced back to this literature largely being built on revealed or stated preference approaches rooted in the use of survey data that are sometimes extended with complementary datasets. However, in few of these previously used datasets there is explicit information on the (potential) spatio-temporal variability in passengers' choices for an airport. Research in this vein has thus produced relevant yet aggregated insights into what drives airports' attractiveness in a MAR. Airport attractiveness in a MAR context may nonetheless be contingent upon the time of day, the day of week, or even exhibit seasonal variations. An obvious example is an airport's accessibility, which is commonly shown to be a key factor in MAR airport choice: due to congestion, the geographies of airport access time by road may well be different at 8AM or at 12PM, and they may also be different on a weekday or on a Sunday. This is corroborated by earlier studies on airport accessibility measured by travel time by car, which has been shown to be fundamental in the choice for an airport (Skinner, 1976; Harvey, 1987). Furthermore, Hess and Polack (2005a, 2006) showed that air passengers envisage increasing travel times to the airport as an increasing risk to miss their flight. Koster et al. (2011) drew on this research to develop a mixed logit model designed to measure the effect of airport access travel

time variability on access travel cost. Their findings substantiate the premise that both business and non-business passengers are sensitive to higher travel time costs when accessing an airport. Furthermore, airport and airline schedules are often synchronized with the organization of a 'working day' with more flights in the early morning and evening (Budd et al., 2011). Taken together, there is now an extensive body of research focusing on uneven car accessibility to airports and its cost, as well as on the scheduling dynamics of airlines in response to demand. However, to the best our knowledge, to date there has been no research on how these complex dynamics affect (catchment areas in) MARs.

One of the possible reasons why previous MAR research has not considered the dynamic nature of catchment areas is the longstanding lack of detailed data on the spatio-temporal dimensions of the accessibility and utility of airports. However, recent developments in data availability and analysis are gradually opening up new opportunities in this field of research. Using the example of domestic connections departing from the New York MAR, in this paper we report on the development of a framework that allows analysing spatio-temporal dynamics in airports' catchment areas. Rather than drawing on revealed or stated preferences as captured by survey data (and therefore actual behaviour), we devise ideal-typical catchment areas by considering some commonly cited airport choice drivers (and therefore supposed behaviour rooted in a notion of utility-maximization). Our purpose, therefore, is not to present a comprehensive analysis of catchment area dynamics in the New York MAR. Rather, we use this setting to develop a flexible framework that allows exploring if, when and how such spatio-temporal dynamics matter. To this end, for a series of different time windows, (1) we analyse the geographies of *access* time by road to the different MAR airports; (2) parameterize MAR airport *utility* based on pricing, connectivity characteristics, and on-time performance; and use this information as the input to (3) a Huff model to calculate different airports' *attractiveness* and associated *catchment areas* at the level of census block groups (US Census Bureau, 2018). Comparing these catchment areas for the different time windows then allows revealing some elements of their spatio-temporal dynamics. Importantly, this framework can be adapted in follow-up research to include other measures and combinations of airport accessibility and utility, and subsequently implemented irrespective of the specific MAR context.

The remainder of this paper is organized in four main sections. First, we further discuss our analysis in the context of the MAR literature. Second, we outline our model specification. Third, we show the ramifications of our approach by discussing some empirical patterns of spatio-temporal dynamics in the New York MAR. In the fourth and final section, we provide an overview of our main findings and reflect on possible avenues for further research.

4.2 Literature review

4.2.1. Airport choice, airport attractiveness, and catchment areas

As there is now an extensive body of literature on MARs, in the below discussion we focus on those elements that are particularly relevant for our paper – reviews of this research can be found in Hess & Polak (2005b), Muñoz et al. (2017), and Fuellhart & O'Connor (2019). Our starting point is that although one of the defining features of a MAR is that passengers have in principle a *choice* between airports, in practice this choice is *constrained* by a number of considerations. Previous analyses of airport choice in MARs have consistently shown that there are two sets of considerations that make an airport more attractive than others: (1) its accessibility and (2) its utility (e.g. Pels et al., 2001; de Luca, 2012). Although this has only rarely been broached, this implies that MAR airport can be studied in the spirit of a Huff model (see, for example, Heilman, 2017): a spatial interaction model that calculates gravity-based probabilities of customers (i.e. passengers in a MAR) at each origin location choosing a facility (i.e. a particular MAR airport) from all potential facilities (i.e. all MAR airports). As probabilities in a Huff model are derived from information on facilities' accessibility and utility, MAR airport choice can be cast in this form. The results of a Huff model can be spatially represented in the form of catchment areas: the geographical area from which a facility attracts (the bulk of) its customers.

However, to date airport choice models and the catchment areas that can be derived from these have most commonly been based on *revealed* (Fuellhart, 2007; Paliska et al., 2016; Heilman, 2017) or – somewhat less often – *stated behaviour* (Loo, 2008). Using a Huff approach, in contrast, entails modelling airport choice based on *assumptions about behaviour*: the way in which accessibility and utility influence choice-making. In this case, this behaviour is expressed in terms of a notion of utility-maximization: choice sets specifying how prospective passengers will likely value, weight and combine (different elements of) airports' accessibility and utility for a trip when multiple reasonable choices exist. Using a Huff model based on assumed rather than revealed or stated behaviour has drawbacks, the most important one being that – especially without validation – the real-world relevance of these models is contingent upon the accuracy of their underlying assumptions. However, airport choice models based on surveys have a number of disadvantages in their own right. First, survey data may be costly, proprietary or absent. Second, MAR research risks to some degree being data-driven in that surveys do not necessarily reflect researchers' interests but rather those of airports and airport planners who are often responsible for the survey organization. Our research focus is a case in point, given that little if any information in the most commonly cited MAR surveys considers the (possible) relevance of the time window to which the choice pertains. And third and finally, MARs are context dependent and surveys to some degree idiosyncratic, which makes comparing and generalizing across MARs difficult (cf. Fuellhart & O'Connor, 2019). Using a model built on the assumed choice for airports that credibly reflects earlier insights may to some degree address these drawbacks. This is the path we follow in this paper. To this end, below we discuss which dimensions and variables have in previous research been repeatedly shown to significantly influence airport choice in MARs.

4.2.2. Main elements of airports' attractiveness in MARs

A first element in airport choice models and a Huff model alike is the accessibility of the facilities. Irrespective of the geographical setting or the type of passenger, airport accessibility always emerges as a key feature of the attractiveness of an airport (Pels et al., 2003; Hess & Polak, 2006; Hess, 2010). Airport market shares are generally high in the areas close to the airport, and decrease when moving further away (Lieshout, 2012). In the New York MAR, for example, it is clear that – all other things being equal – LaGuardia Airport (LGA) would emerge as the most reasonable option when departing from northern Brooklyn. Several methods have been proposed to measure airport accessibility. Sometimes this straightforwardly entails specifying a general radius around the facility of interest (MarcUci & Gatta, 2011) or drawing Thiessen polygons (Zhou et al., 2018). More refined approaches focus on the driving distance to an airport (Fuellhart, 2007), consider access time (Zhou et al., 2018), incorporate modal alternatives (Paliska et al., 2016), take a cost perspective (Koster et al., 2011), or advocate adopting a

broader perspective on access time by also considering potential time-savings associated with using smaller or more efficiently organized airports (Distill, 2013).

In our analysis, we measure accessibility as airport access time by road. Whereas in Europe and other parts of the world airport-city rail links tend to be fairly well adopted, in the US ridership is estimated to be only between 2% and 10% (International Air Rail Organisation & Blond, 2013). Bearing this in mind, and given the limited number of airports serviced by rail connections in the US (Goetz & Vowles, 2009), we focus exclusively on road links to each airport. Ishii et al. (2009) showed that passengers are very sensitive to the time costs associated with airport accessibility in a MAR. They found that even small differences in access time (e.g. even a 5-minute reduction) are enough to lead to noticeable shifts in airport choice. The relevance of this measure is corroborated by Loo (2008), who found that access time was a statistically significant variable in MAR airport choice, whereas the number of access modes, access costs, and queue time at check-in counters were not. One further advantage of using access time by road is that it arguably allows for a direct illustration and measurement of a major dimension of spatio-temporal variability in catchment areas: detailed traffic data allows to better show how airport accessibility varies in space and time, especially in the face of geographically complex congestion patterns (Gallotti et al., 2017).

The second element in airport choice models and a Huff model alike is the utility of the facilities themselves. For example, LGA may well be the closest facility to Northern Brooklyn, but it is also a de facto national airport which would make it less-than-useful for passengers flying internationally. Measurements of airport utility are bound to be more complex than measurements of airport accessibility because they involve a broad range of interlocking variables. For example, a survey of the literature suggests that *inter alia* flight frequencies, the presence of direct connections and/or the number of stops to reach a destination, airfare, type of aircraft, purpose of travel, socio-economic considerations, loyalty programs offered by the airlines serving an airport, the number of passengers traveling together, previous consumer experiences, and a range of other variables may wield an influence on a passenger's choice. In addition to there being many possible variables, there is also no consensus about which of them matter the most, how they interact, and how much of this depends on the particular MAR context. Below, we single out three choice sets that often resurface in MAR analyses and which will be operationalized in a series of utility variables.

First, the impact of air fares on MAR airport choice has been repeatedly shown. With the emergence of low-cost carriers (LCCs), the landscape of air travel has considerably changed (Windle & Dresner, 1995; Franke, 2004; Pitfield, 2008; Cho et al., 2015). When choosing a carrier or an airport, air fares are often more important than the services being offered in return (Blackstone et al., 2006; Pitfield, 2008; Fuellhart et al., 2013). Zhang & Xie (2005) found that 60% of leisure passengers and 45% of business passengers rated ticket fare as the most important factor when choosing a flight. The importance of air fares on catchment areas has also been shown by passengers willing to travel further and/or longer in exchange for a better air fare (Dresner et al., 1996b; Suzuki & Audino, 2003; T. H. Grubesic & Matisziw, 2011).
Second, there are a number of connectivity characteristics that may influence airport choice. There seems to be a consensus that passengers prefer non-stop or fewer-stop routes (Vowles, 2001; Bounova, 2009; Hsiao & Hansen, 2011; Mandel, 2014; Park & O'Kelly, 2016), while markets served and flight frequency have also been revealed to be determinants of airport choice (Pels et al., 2000, 2003; Fuellhart, 2007; Fuellhart et al., 2013).
A third and final dimension is the on-time performance of airports. The literature suggests that even though passengers rarely state they consider on-time performance when selecting an airline or airport (Suzuki, 2000; Hess, Adler, & Polak, 2007; Hess & Polak, 2005a), the rate of passengers switching to a new airline or airport has been shown to be considerably higher for those with recurring delay or cancellation experiences (Suzuki, 2000). Ishii et al. (2009) found that the marginal effects of access time on the choice probability of a flight is larger than those of possible delays (which they attribute to

informational constraints), but this nonetheless confirms that delays are also often factored into the choice process.

Airport choice in a MAR is obviously more complex than outlined in the above overview. Individual passengers' previous experiences with an airport may matter. Ishii et al. (2009), for example, found that, after controlling for accessibility and a range of utility variables, travellers in the San Francisco-Los Angeles market still had residual airport preferences for San Francisco International Airport (SFO) and Los Angeles International Airport (LAX). Furthermore, different types of travellers will value utilities differently. The MAR literature commonly distinguishes between business and leisure travellers (even though that distinction between both is more complex than often acknowledged, as shown by Lassen (2006)), with the former often shown to be less sensitive to air fare and more sensitive to delays. Ishii et al. (2009) also argued that passenger choice may involve a joint airport-and-airline decision, with – all other things being equal – airline brand loyalty playing a role in a decision for an airport. Taken together, then, neither the operationalization of accessibility (access time by road) nor the choice sets and the variables therein (fare, connectivity characteristics, on-time performance) used in our Huff model paint an exhaustive picture of MAR airport choice, with above all individual travel characteristics possibly playing a supplementary role. However, as these cannot be easily modelled at an aggregate spatial level (see, however, Fuellhart, 2007), they are not considered in our framework. That said, we believe it provides us with a framework that captures the most commonly observed effects that tend to have the largest effects. For example, Ishii et al.'s (2009) analysis attributes the largest marginal effects to the building blocks of our Huff model: alongside accessibility, they mention lower air fares, higher frequencies and shorter delays.

4.2.3. Spatio-temporal variability in catchment areas

The key purpose of our Huff model is not to model catchment areas per se, but to use it as a lens through which we can explore some of the spatio-temporal variability in these catchment areas. Both airport accessibility and the different utility choice sets clearly display multiple patterns of spatio-temporal variability. In an analysis of taxi drivers' airport pickup decisions in New York City, Yazici et al. (2013) showed that, in addition to airport proximity and operational variables such as driver shift organization, the time of day had a significant impact on their willingness to serve JFK. Based on their findings, they argue that adding more flights to JFK around rush hour would not necessarily translate into more passengers choosing the airport: during this time window, improved airport utility because of a larger number of flights would be partially offset by a reduced airport accessibility as most roads to access JFK from New York City are highly congested at that time of day. Airport utility, in turn, can also vary depending on the time window. Harvey (1987) already linked airport choice to scheduling factors, especially in the case of sufficient supply/demand of/for specific connections. An obvious example in the New York MAR would be that LGA has a perimeter rule that limits nonstop flights from and to the airport beyond a 1500 miles radius, with the exception of connections to Denver and on Saturdays. In addition, some flights are only organized seasonally. For example, during the period this research was carried out, Delta Airlines had a seasonal Saturday-only direct flight from LGA to Bozeman Yellowstone International Airport (BZN). Thus, LGA's utility may be higher and therefore its catchment area larger on a Saturday in summer. In addition to scheduling, on-time performance may also vary depending on the time of day. For example, slot constraints at JFK and EWR have been implemented to curb delays during peak times, but these having hardly been effective (Luttmann, 2019). The net result is that airports' overall attractiveness may vary depending on the time window: choosing a 9AM flight from JFK may both incur excess airport access time and a higher chance of the flight itself being delayed; both its accessibility and (this dimension of) its utility may nosedive during this particular time window, and its catchment area may thus shrink depending on the situation at other MAR airports.

The above examples are of course both idiosyncratic and anecdotal, but the key implication is that MAR airport catchment areas may exhibit spatio-temporal variability. To assess this variability,

researchers need to have access to data sources that explicitly allow them to evaluate how airport accessibility and utility evolve throughout the day, week, or even season. A number of developments that are commonly associated with the 'big data revolution' (Kitchin, 2014) have opened up new possibilities in this regard (Sun et al., 2017). In terms of accessibility, Gallotti et al. (2017) point out that recent evolutions in and the popularization of the use of Information and Communication Technologies (ICT) now provide data sources that allow for a detailed assessment of the accessibility of airports[63]. In terms of utility, in turn, there have recently been efforts to automate the generation of bespoke datasets from publicly available information sources. Teixeira and Derudder (2018) developed an open-source software package that allows generating tailored air transport data from the datasets provided by the Bureau of Transport Statistics (BTS) in the United States, including parsing data for different time windows. In our framework, outlined in the next section, we make use of these developments and data sources to specify our Huff model.

[63] In their paper, Gallotti et al. (2017) investigate how the availability of ICT data such as GPS records of taxi pickups, geolocated tweets, and travel time made available via Google allow for new and more accurate depictions of travel behaviour in general and airport access in particular.

4.3 Analytical framework

4.3.1. The New York MAR

We apply our framework to the New York MAR. In principle, a MAR can be formally defined as a set of two or more airports that commercially serve a single regional market (de Neufville, 1995). However, in order to be useful, this 'single regional market' needs a formal definition. It is in principle possible to identify MARs based on IATA's specification of regional codes. Thus, IATA's NYC code is used to collectively refer to John F. Kennedy International Airport (JFK), LaGuardia Airport (LGA), Newark Liberty International Airport (EWR), and Stewart International Airport (SWF). Unfortunately, this approach cannot be systematically used to identify MARs as some clear-cut examples do not have an IATA code[64]. More importantly, however, many of these designations are intuitive notions that are not based on an analysis of whether and how the regional market functions as an integrated, functional whole.

Recent literature has therefore developed functional approaches to identify MAR configurations. Brueckner et al. (2014), for example, marshal a methodology based on incremental competition effects from nearby airports on average fares in a MAR's primary airport. Their results for the United States corroborate the relevance of research into MAR identification in general and the interaction between MAR airports in particular, as there is evidence that MAR-pairs rather than airport-pairs often provide a more appropriate market definition for analyses of air passenger transport. Here we follow their description of the New York MAR, which is based on a two-step approach. Their first step was the identification of the metro area's 'primary' airport, i.e. the airport that served the largest number of domestic origin and destination passengers (LGA). The second step in their categorization consists of identifying 'core' and 'fringe' airports based on distance to that primary airport (which is also part of the core category). In the case of the New York MAR, this led – compared with IATA's NYC code – to the exclusion of Stewart International Airport (SWF) and the inclusion of Westchester County Airport (HPN) and Long Island MacArthur (ISP). The potential catchment area associated with this MAR is defined as the New York Metropolitan Area, which includes New York City, Long Island, and a selection of other proximate parts of the states of New York, New Jersey, and Connecticut. For the sake of simplicity, in the remainder of this paper these five airports and the New York Metropolitan Area represent our working definition of the 'New York MAR'. The location of the airports and the catchment area are shown in Figure 16. This represents an area of approximately 57000 km^2 and a population of approximately 30 million by 2017 Census estimates (US Census Bureau, 2017).

[64] For example, even though San Francisco International Airport (SFO), Oakland International (OAK), and Norman Y. Mineta San Jose International (SJC) are clearly part of a single MAR (e.g. Harvey, 1987; Pels et al., 2011), there is no single IATA code bringing together There is a broader set of MAR-like airport codes that is used across airline booking systems, listed on the Wikivoyage website (https://en.wikivoyage.org/wiki/Metropolitan_area_airport_codes). In this case, there is a code for the Bay Area (QSF) covering SFO, OAK and SJC. However, such a 'metropolitan area airport code' is – unlike WAS and NYC – not an actual IATA code, but rather a convenient shorthand used in some airline booking systems. This implies that, say, QSF will not work for all bookings systems, while different booking systems may refer to different regional/metropolitan realities (e.g. the uneven inclusion of Westchester County Airport (HPN) in the NYC code).

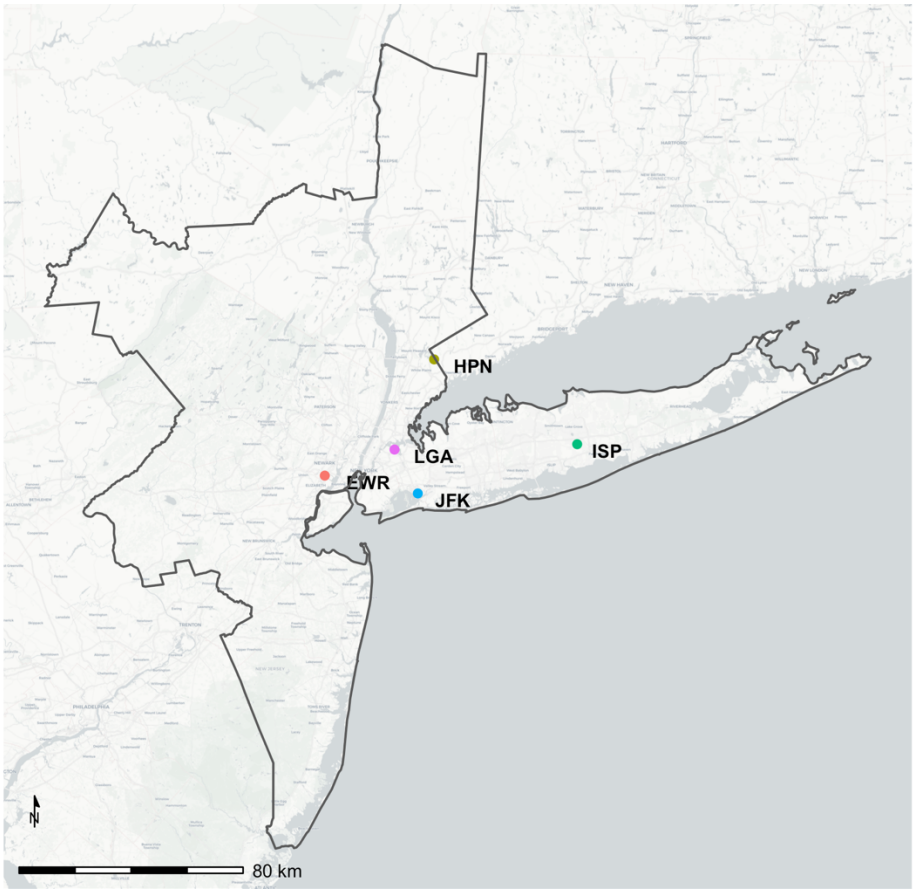


Figure 16 - New York MAR airports and potential catchment area (New York Metropolitan Area).

Airport Name	IATA code	Passengers (thousands)	Departures (thousands)	Available Seats (thousands)
Newark Liberty International	EWR	15922	157	18948
Westchester County	HPN	775	13	960
Long Island MacArthur	ISP	822	6	1046
John F. Kennedy International	JFK	14051	128	16797
LaGuardia	LGA	13963	164	17424

Table 9 – Number of passengers, departures and available seats per airport in the New York MAR, 2018 (Bureau of Transport Statistics, 2020). *Only domestic flights are included.*

Table 9 provides some basic statistics about these five airports. Based on the number of passengers, departures and available seats, it is obvious that EWR, JFK and LGA stand out in terms of overall connectivity. There are nonetheless some differences in the nature of that connectivity, most clearly epitomized by the earlier-mentioned perimeter rule at LGA. A further obvious difference is that LGA is therefore also almost exclusively a domestic airport[65]. However, this lack of international connections is less relevant here, as our empirical focus is only on the choice for, and the catchment areas associated with domestic flights.

[65] There are some international connections, for example between LGA and Toronto Pearson International (YYZ), because the presence of US Customs and Border Protection at YYZ enables travellers to the US to clear customs and immigration before boarding their flight.

4.3.2. Modelling approach

Because facilities rarely have explicit catchment areas with fixed and impermeable boundaries, the best way to model them has been extensively debated in the geographical literature (cf. Harris et al., 2016). The catchment areas produced by a Huff model are shaped by the attractiveness P_{ij} for a customer (passenger) located in geographical area j when choosing facility (airport) i. Catchment areas can be formulated from the perspective of the facilities (airports), showing where they source the majority of their customers (passengers). However, it is also possible to formulate catchment areas from the perspective of the geographical areas, showing which facility (airport) the largest proportion of customers (passengers) living in the area will choose. In our results, we combine both approaches by showing which airport has the 'best' combination of accessibility and utility for each location and how much this particular choice 'prevails' over others.

When casting MAR airport choice along the lines of the classical formulation of a Huff model (Huff, 1963, 1964), P_{ij} is given by:

$$P_{ij} = \frac{{U_i}/{D_{ij}^{\alpha}}}{\Sigma_{i=1}^{n}\left({U_i}/{D_{ij}^{\alpha}}\right)} \qquad (1)$$

With:
P_{ij}: the attractiveness of airport i for passenger departing from geographical area j;
U_j: the utility of airport i;
D_{ij}: the distance from geographical area j to airport i;
α = an exponent applied to distance so that the attractiveness for distant airports is reduced.

In our model, distance D_{ij} is operationalized as driving time by road; utility U_i is operationalized for three choice sets separately and conjointly (fare, connectivity characteristics, on-time performance); and α is set to 1 as there is no specific reason to presuppose a non-linear relationship. This implies that we calculate four different sets of catchment areas. The geographical areas j are census block groups: the smallest geographical unit for which the US Census Bureau (2018) publishes sample data, and which typically have a population between 600 and 3000 people. The airports i are in principle the five airports specified in Brueckner et al. (2013), but note that when calculating P_{ij} we only consider those airports that are within a 60-minute driving time (see below).

The operational implementation of the four versions of the Huff model – one for each utility choice set and one for overall utility – involves three consecutive steps whose operationalization will be elaborated in the remainder of this section:

(1) For each census block j, we determine their driving time by road to the airports (D_{ij});
(2) For each airport i, we determine their utilities (U_i);
(3) For each census block j, we calculate the attractiveness P_{ij} for a passenger departing from this census block when choosing from the different airports i that are within a 60-minute reach based on Equation 1.

Maps are then drawn that show, for each census block group, which airport would be the best choice assuming that passengers will value, weight and combine accessibility and choice sets as per our operationalization of the model. As mentioned, given the main objective of our paper – assessing spatio-temporal variability in MAR airport catchment areas – the drawback of working with assumptions is of secondary importance: using a modelling approach allows us to research the effects of different time windows on catchment areas. To this end, we implement the Huff models for different time windows: for four different time windows throughout the day (7am-10am (peak am), 10am-4pm

(midday), 4pm-7pm (peak pm), 7pm-12am (evening))[66]; for the different days of the week; and for the different quarters of the year. Again, different sets of implementations can easily be produced depending on the research questions at hand (e.g. weekdays versus weekend, monthly time windows, different time windows throughout the day).

4.3.3. Operationalization of variables

We use driving time by road to each of the selected airports as our measure of D_{ij}. For the four different time periods throughout the day, we use a central time (i.e. 8.30am, 12pm, 5pm, 8pm). The traffic data was supplied by Here Maps (2020)[67], through the ArcGIS Online traffic plugin (ArcGIS, 2019) and parsed in R (R Team, 2017). This data gives, for each census block group, the access time to an airport in consecutive bands (20 mins, 30 mins, 45 mins, 60 mins). Although this does not allow for a precise specification of access time, it does allow for a reasonable estimate of variability throughout the day. It also explains why, when implementing the Huff model, we only consider those airports that are with a maximum driving radius of 60 mins. The inclusion of a 60 minutes driving radius allows us to focus exclusively on the NY MAR. Increasing the driving radius to 150 minutes would imply that other factors (e.g. airport leakage) would influence the operational setup of this paper, and steer the focus away from our methodological objective (Fuellhart, 2007; Lian & Rønnevik, 2011). By way of example, Figure 17 shows the daily dynamics in driving time to JFK on a Monday by means of these access time zones. It clearly shows the airport being more difficult to access at 8.30 AM and especially 5PM, with almost no access below 60 minutes possible from west of the Hudson River. At 8PM, however, JFK can be reached in less than 60 minutes from a fairly large number of census blocks located west of the Hudson River. Based on these data, Figure 18 shows the capacity of each of the airport to capture potential passengers (cf. Frost & Spence, 1995; Gutiérrez, 2001; Wu, 2011) from within the New York Metropolitan Area for the different time windows on a Monday, Saturday and Sunday. Because of their central location viz. the population distribution, LGA, EWR and JFK unsurprisingly have the potential to capture more passengers than HPN and above all peripheral ISP. However, the key point here is the spatio-temporal variability in accessibility shown in Figure 18. Temporal variability is for example shown from peak AM and peak PM on a Monday generally being the time window during which airports capture the fewest passengers. LGA is able to capture up to 34% more passengers in the evening than during the peak PM period, which represents about 4 million extra potential passengers. Patterns are different over the weekend, where the peak AM time window is the least congested time of day. Importantly, this temporal variability is cross-cut by spatial variability. ISP, for example, shows less variability than the other airports, while LGA clearly suffers more than EWR during Monday peak PM from a shrinking potential catchment area.

[66] This leaves us with a time window between midnight and 7AM, but given low number of flights the utility data would have been too sparse to draw up conclusions.

[67] ArcGIS traffic data is based on the average of observed speeds over the past three years.

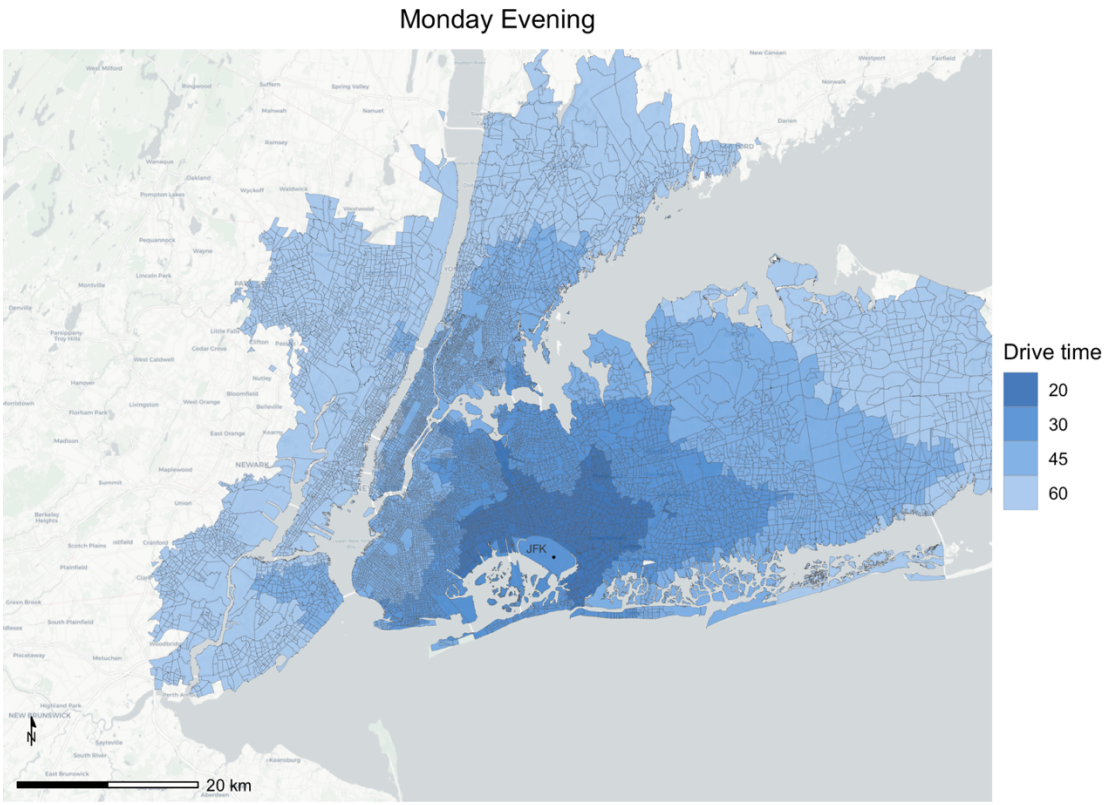


Figure 17 - Driving times to JFK airport on a Monday evening at the level of census block groups.

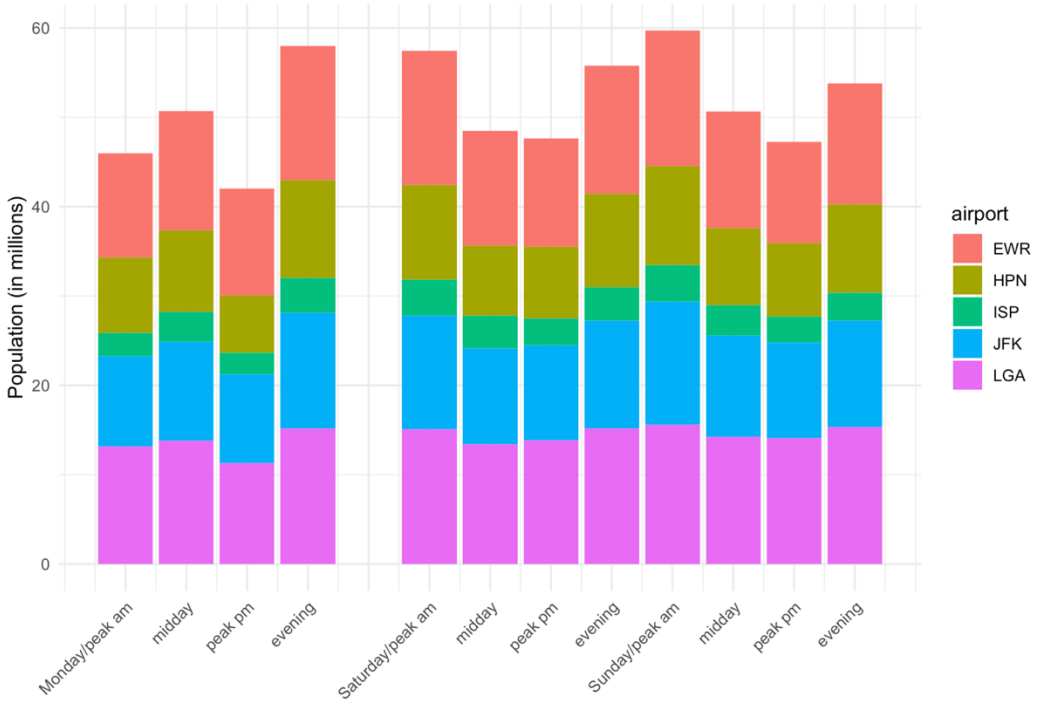


Figure 18 - Population potentially captured by each airport (i.e. aggregated population of all census block groups within 60 minutes from an airport) for 2018 (averaged Q1 throughout Q4).

To operationalize the three different utility choice sets, we use domestic travel data retrieved from the US Bureau of Transport Statistics (BTS), more specifically their Origin Destination Survey (DB1B), Air

Carrier Statistics (T-100) and Ontime Performance databases. The data respective to 2018 were processed with the R SKYNET package (Teixeira & Derudder, 2018). Following Lieshout (2012), for each airport we first identified their market area as the destinations that are reachable with a maximum of two legs. All choice sets and their constituent variables were calculated for the different time windows, with the exception of the fare variable which is static as it was impossible to match fares to time windows (i.e. analyses of fare attractiveness will only vary by access time). As will become clear, in each step of the process we apply min-max normalizations so that the lowest utility equals 0 and the highest utility equals 1. This is done for ease of interpretation, the possibility of straightforwardly combining different variables, and reasons of comparability across time windows.

The air fare utility choice set Uf_i is operationalized by attributing a score of one to the airport offering the lowest average fare to a market, and aggregating these scores across all markets served:

$$Uf_i = \sum_m Cheapest\ MAR\ fare_m \qquad (2)$$

With

- Uf_i : air fare utility of airport i;
- Cheapest MAR $fare_m$ = 1 if airport i offers the cheapest average price across MAR airports to market m, and = 0 if otherwise[68].

For each time window, the calculation of all air fare utilities Uf_i is followed by a min-max normalization.

The connectivity characteristics utility choice set Uc_i is based on a combination of four variables: number of markets served, number of departures, number of most directly served markets, and number of unique markets. We first apply a min-max normalization to each of these variables across airports, after which we combine them into a connectivity characteristics utility Uc_i as follows:

$$Uc_i = \#Mk_i + \#Dp_i + \#Md_i + \#Mu_i \qquad (3)$$

With

- Uc_i : connectivity characteristics utility of airport i;
- Mk_i : Markets served by airport i;
- Dp_i : Departures at airport i;
- Md_i : Most directly served markets by airport i;
- Mu_i : Unique markets served by airport i.

For each time window, the calculation of connectivity characteristics utilities Uc_i is followed by a min-max normalization.

The on-time utility choice set Uot_i combines on-time, delayed, and cancelled flights. Some of the models using operationalizations of on-time flights are based on what appear to be arbitrary operational definitions (e.g. Hess et al., 2007; Suzuki & Audino, 2003). Here we adopt the approach and data of the US Department of Transportation (DOT), which considers a flight to be on-time if arriving or departing within 15 minutes of scheduled time (US Department of Transportation, 2019). This approach has been used in earlier studies (e.g. Dresner & Xu, 1995; Suzuki & Tyworth, 1998; Pels et al., 2001), and here we extend it by also considering cancelled flights. The total number of flights are disaggregated into cancelled, delayed and on-time flights, and considered in relation to the total number of flights. To calculate our overall measure Uot_i of on-time utility, we aggregate the relative number of cancelled, delayed and on time flights with the former two incurring a penalty by means of an exponent of -1:

[68] Note that this implies that an airport de facto scores a 1 for unique markets.

$$cancelled = \sum_{m} min\ cancelled\ flights\ MAR_{m}$$

$$delayed = \sum_{m} min\ delayed\ flights\ MAR_{m}$$

$$ontime = \sum_{m} max\ ontime\ flights\ MAR_{m}$$

$$Uot_i = cancelled + delayed + ontime \qquad (4)$$

With

- Uot_i : on-time utility of airport i.
- min cancelled $flights_m$: 1 airport i has the lowest percentage of cancelled flights across MAR airports to market m, and = 0 if otherwise
- min delayed $flights_m$: 1 airport i has the lowest percentage of delayed flights across MAR airports to market m, and = 0 if otherwise
- max on-time $flights_m$: 1 airport i has the highest percentage of on-time flights across MAR airports to market m, and = 0 if otherwise

For each time window, the calculation of all on-time utilities Uot_i is followed by a min-max normalization.

And finally, we create an overall utility measure $Uscore_i$ by aggregating the three utility sets to be used:

$$Uscore_i = Uf_i + Uc_i + Uot_i \qquad (5)$$

It is possible to sum and include in the $Uscore_i$ only the relevant utilities (i.e. connectivity, on-time, fare) depending on the research question. For each time window, the calculation of all total utilities $Uscore_i$ is followed by a min-max normalization. This then results, for each time window, in four different utility choice sets – Uf_i, Uc_i, Uot_i, and $Uscore_i$ – that can be used together with driving time measures D_{ij} as the input to the Huff model given by Equation 1. This produces, for each census block j, the attractiveness P_{ij} for a passenger departing from this census block to choose for the different airports i that are within a 60-minute radius. Our model also allows for specific research questions based on specific destinations to be made. For example, if the interest of the study is passengers flying to destination k, airport i is selected if it offers flights to that same destination.

4.4 Results: spatio-temporal dynamics in the New York MAR

Given that there are many possible combinations of time windows and utility choice sets, in this section we will restrict ourselves to a limited set of illustrative examples that collectively show the potential of the approach. Results are primarily conveyed trough catchment area maps. Figure 19, for example, shows the catchment areas associated with the overall utility choice set $Uscore_i$ for the four time windows on a typical Monday. In this and the subsequent maps, each census block group j assumes the colour of the airport i for which it has the largest attractiveness P_{ij}, with lower values of P_{ij} resulting in a more transparent shading. Note that the maps may be slightly deceptive when it comes to estimating the total (market) size of the catchment areas, as census block groups in the western, eastern, and northern fringes generally cover much larger areal surfaces than those in, say, Manhattan. To corroborate that catchment areas are, generally speaking, commensurate with the ratios shown in Table 1, Figure 20 shows the total number of individuals 'captured' by an airport's catchment area for different time windows in proportion to the P_{ij}. We will return to some of the patterns in this figure below, but for now the point is that the overall lower attraction of ISP and HPN is confirmed here even though their catchment areas appear to cover quite substantial parts of the New York Metropolitan Area. EWR, in turn, exceeds LGA and JFK, which can be explained by the latter two airports being more in competition given their location: very few census block groups are exclusively in the 60 minutes realm of one of these airports, while a fairly large number of block groups west of EWR are.

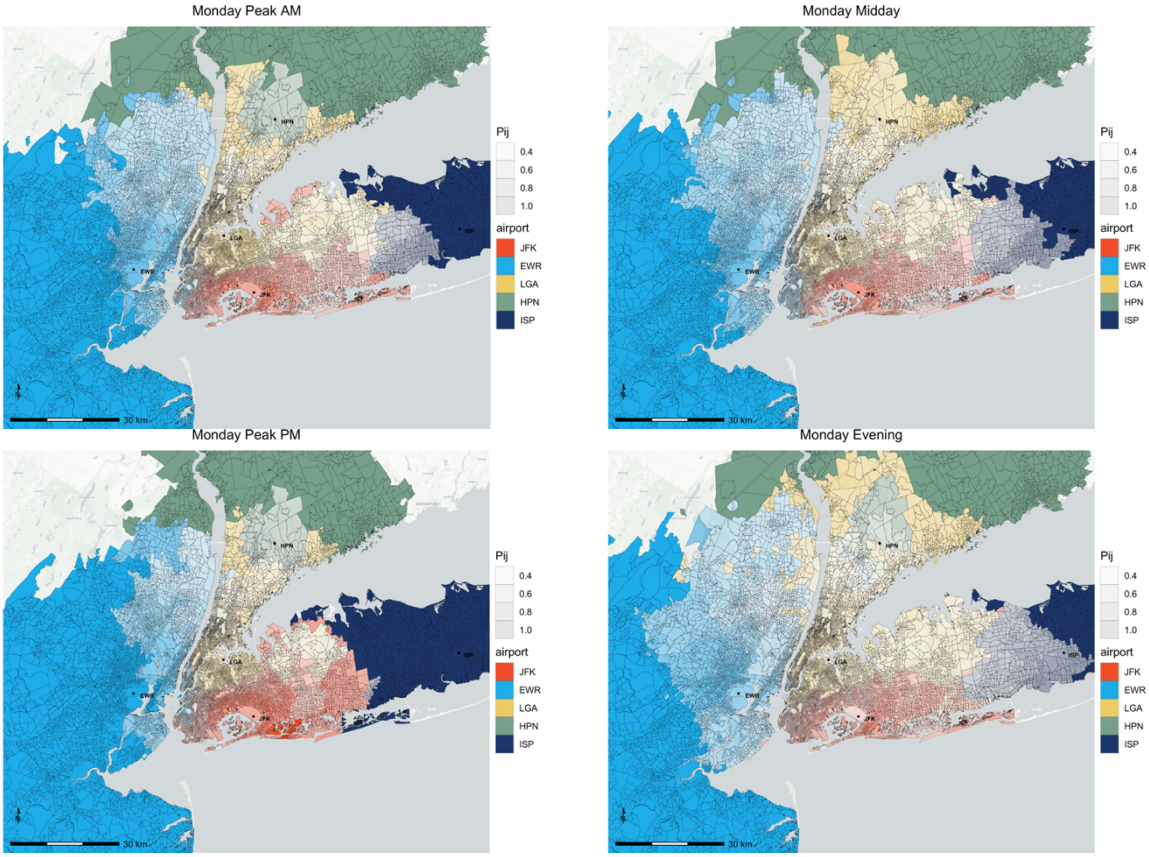


Figure 19 - Catchment areas associated with the overall utility choice set $Uscore_i$ for the four time windows on a typical Monday.

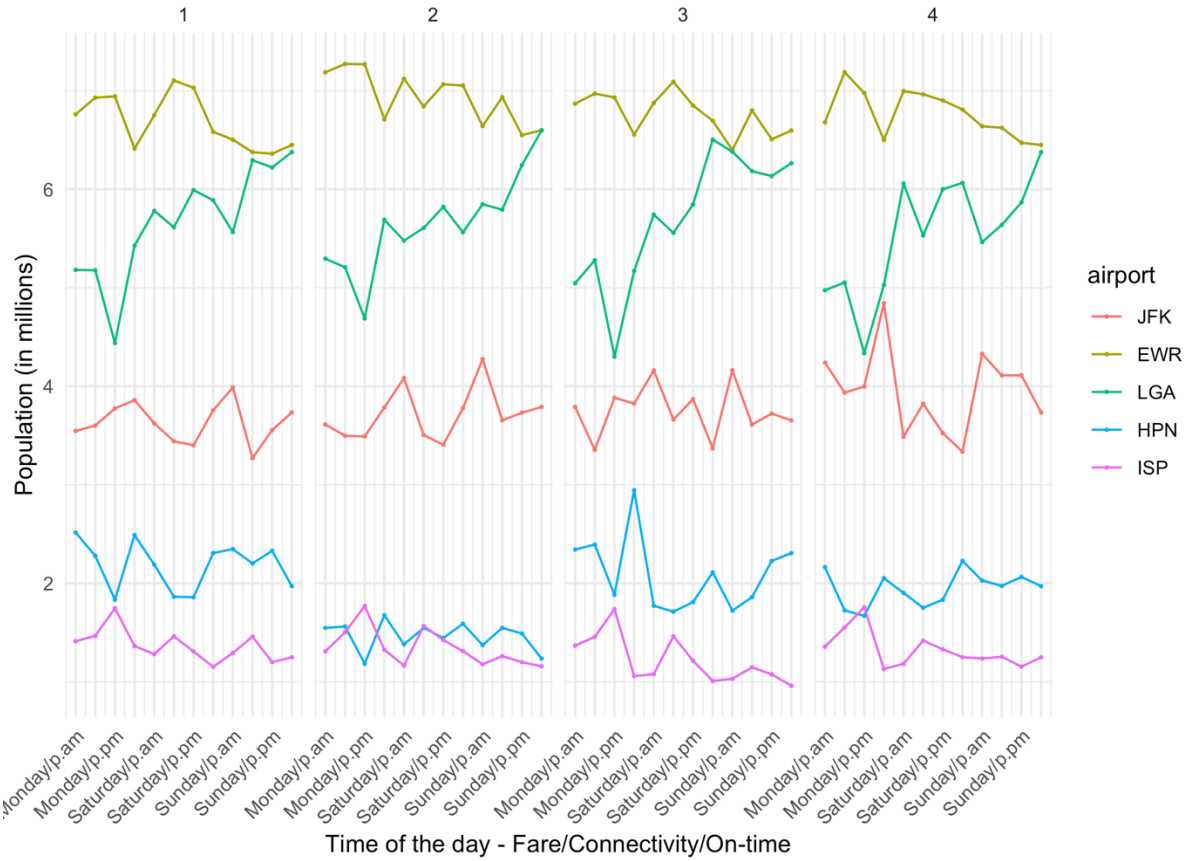


Figure 20 - Catchment area sizes associated with the overall utility choice set Uscore$_i$ for the four time windows, different days of the week, for Q1 through Q4.

Figure 19 and subsequent catchment area maps unsurprisingly show that census block groups that are either very close to an airport and/or imply travel close to an airport *en route* to the other airports tend to be straightforwardly assigned to a catchment area. This is shown both by the quasi-consistent assignment of some of the census block groups to the same airport and their generally darker colours. For example, census block groups west of EWR and in the central parts of Long Island tend to be allocated to EWR and ISP, respectively, and this irrespective of the time window or utility choice set. In spite of the stability at the fringes of the NY Metropolitan Area and near the airports, there are however ample catchment area dynamics as well. Most of these are found in the census block groups located 'in-between' two or more airports, such as the Southwestern parts of Brooklyn (i.e. JFK area) or North-eastern parts of Queens (i.e. LGA area). These census block groups tend to have much lighter colours, showing that the choice is more intricate. However, above all, these census block groups sometimes switch colours depending on the time window and/or the choice set, which shows that the Huff model predicts that the 'best' choice is contingent upon these. In the below discussion of results, we will zoom in on these 'in-between' areas, as these are the places where MAR dynamics are most pertinent in geographical terms.

The first example in Figure 19 shows the combination of the total utility Uscore$_i$ and different time windows on a Monday. One of the patterns emerging here is that although EWR regularly emerges as the best choice for the large parts of Manhattan and southwestern Brooklyn, this is no longer the case during the evening period: the driving times and/or the combined choice sets imply that during that particular time window it is not the most feasible choice for census block groups in these areas, with above all LGA and JFK gaining ground. During this period, EWR is also less dominant around the New Jersey/New York state border west of the Governor Cuomo Bridge, with LGA regularly being the best choice. In addition, during the evening, EWR is less dominant in many census block groups north of

the vicinity of the airport. This is also apparent from Figure 20, with the overall size of EWR's catchment area generally falling during the fourth time window (irrespective of the season) and JFK and LGA gaining ground. JFK and above all LGA, in turn, see sharp decreases in the number of passengers they capture during the Monday peak PM time window.
In geographical terms, during the peak PM time window, JFK above all loses ground to ISP. At the same time, however, it is also the time of day when many of the census block groups that are assigned to JFK are firmly within its catchment area (as shown by the darker shadings). In other words: during peak PM, JFK is the best choice for fewer census block groups, but when it is the best choice it is so by far. ISP's catchment area is at it is largest during the peak windows (both AM and PM), which can in part be traced back to Figure 18 showing that it 'suffers' relatively less from congestion. Outside these time windows, ISP's catchment area is smaller with both LGA (in the north) and JFK (in the south) encroaching on its catchment area. This is evident from both the smaller number of census block groups assigned to ISP and those census block groups that are assigned to it being less exclusively associated with it.

The second example in Figure 21 depicts a very different dimension of variability, i.e. at the level of the choice set. In this case, we show the catchment areas associated with the different utilities Uf, Uc and Uot for the Monday peak AM time window. For reference, we also show the total utility $Uscore_i$ which is the linear combination of these three patterns. The expected differences between on-time and connectivity characteristic utilities between HPN/ISP on the one hand and EWR/LGA/JFK on the other hand are obvious. The lower number and the reduced diversity of flights imply that, as far as connectivity characteristics are concerned, the catchment areas of HPN and ISP are basically reduced to the census block groups that cannot reach another airport within 60 mins. It is even likely that, if data availability would have allowed to differentiate beyond the 60 min limit, many of these remaining census block groups would also have been (partially) assigned to one of the three major airports for this particular utility.
At the same time, the limited connectivity of HPN/SIP goes hand in hand with a much better on-time performance, and passengers valuing this utility are therefore often better off at both airports: in this case the HPN and ISP catchment areas extend well south and west, respectively. Although Luttmann (2019) points out that slot constraints at JFK and EWR have been ineffective to curb delays during peak times, it is above all LGA that proportionally suffers the most from an erratic on-time performance, causing its catchment area to dramatically shrink for this particular utility. Results for fare are more mixed and complex, with above all EWR's catchment area shrinking. ISP again makes major inroads into LGA's/JFK's catchment areas, but in this case, losses are compensated by both airports encroaching upon EWR's catchment area. For example, LGA emerges as the best option for fare-conscious travellers in many of the census block groups west of the Hudson near the New York/New Jersey state border.

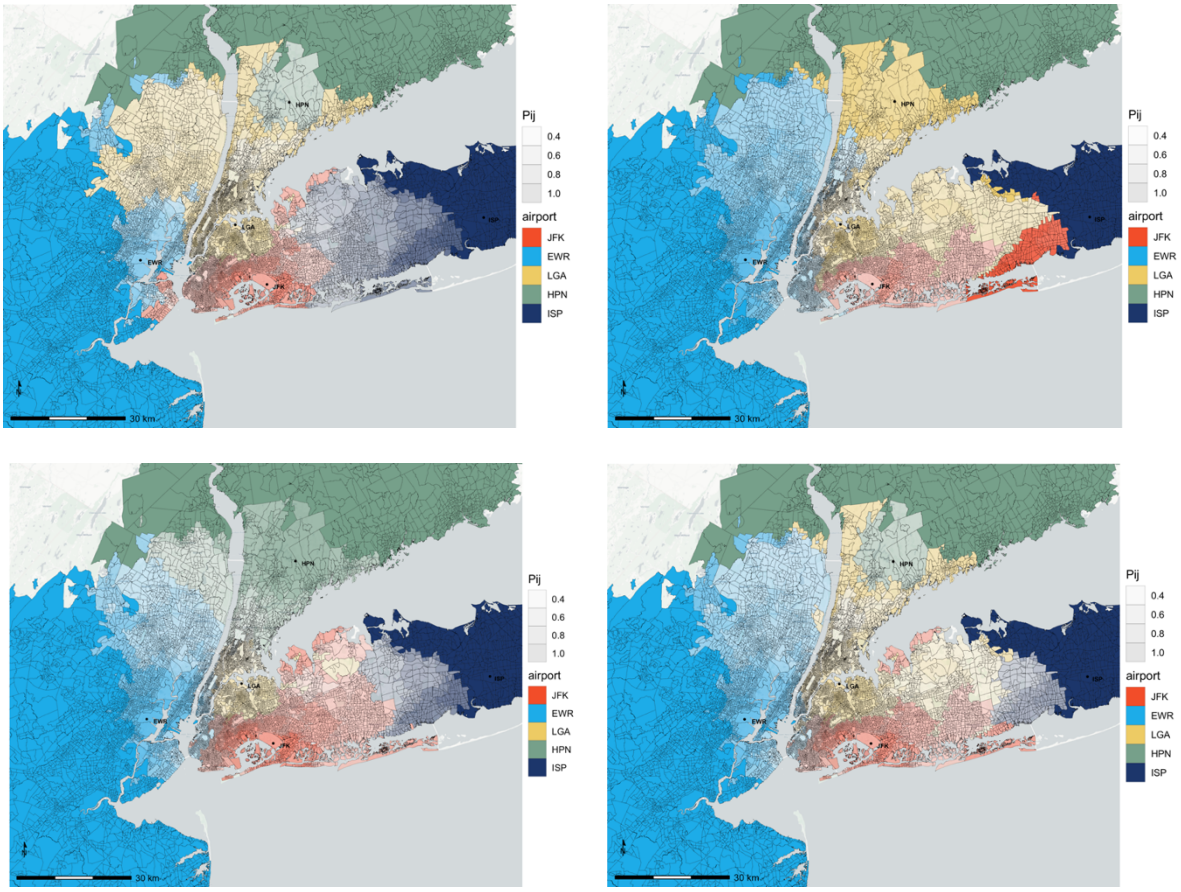


Figure 21 - Catchment areas associated with the fare (top left), connectivity (top right) characteristics, on-time (bottom left) and aggregated utility (bottom right) choice sets during the Monday peak AM time window.

In the third example we go back to comparing different time windows, but in this case, we compare days of the week rather than times of the day. We also zoom in on a specific choice set, in this case connectivity characteristics. Figure 22, therefore, shows the catchment areas associated with the airports' connectivity characteristics during the midday time window on Monday, Saturday, and Sunday. Perhaps the first thing to note here is that LGA's perimeter rule, which does not apply on Saturdays, does not translate into an extension of its connectivity characteristics' catchment area: in theory carriers could use LGA on Saturdays to offer more elaborate connectivity compared to a Monday or Sunday, but this does not appear to be the case.

It is above all LGA and EWR that offer extensive (domestic) connectivity, even resulting in both airports often being the best choice – for this specific utility choice set – for block groups close to JFK. At the same time, there are a number of block groups further east that are outside the 60 min driving distance to EWR and LGA and which are straightforwardly assigned to JFK, and hence the LGA-filled 'void' between the block groups very close to JFK and these block groups further to the east. This void is above all produced by the overall impact of the driving time limit of 60 mins. This also shows from the catchment areas of ISP and above all HPN, which only consist of those block groups from which one cannot reach any other airport within 60 minutes. Thus, ironically: when living next to HPN, LGA is the best available option; but once you move further north, at some point the distance to LGA becomes too large – at least according to our particular model specification – to consider LGA as a viable option, and then HPN emerges as the best option. Of course, if research shows that immediate proximity to an airport plays a major role in the decision process, then the model could be tweaked by changing the nature of the denominator so that block groups next to HPN are effectively assigned to the airport.

The lack of dynamics in LGA's catchment area is reflective of overall limited spatiotemporal variability compared to Figure 19. Nonetheless, there is variability here too, with EWR having a more extensive catchment area – more block groups and generally larger values of P_{ij} – on Mondays.

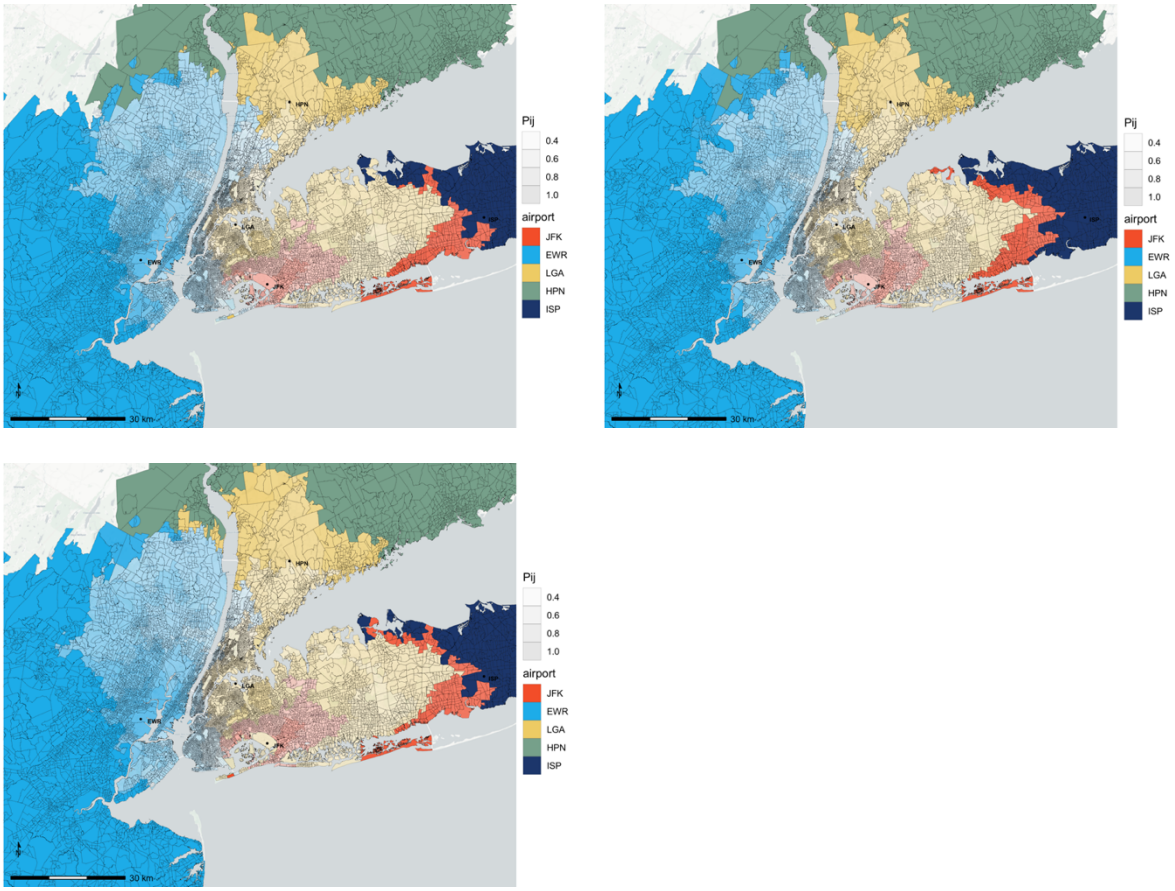


Figure 22 - Catchment areas associated with connectivity characteristics during midday time window on Monday (top left), Saturday (top right), and Sunday (bottom left).

The fourth and final example zooms in on a very different temporal dimension: the different quarters of the year. Figure 23 shows, for each quarter, the catchment areas associated with the fare utility choice set during the midday time window on Sundays. The catchment areas of HPN and ISP are larger during the first quarter, with HPN for example capturing some of the block groups directly next to the airport in competition with LGA. In the three other quarters of the year, the more common patterns of the catchment area of HPN only consisting of those block groups from which one cannot reach any other airport within 60 minutes re-emerges.

The maps also suggest that LGA's catchment area is at its smallest in the 4th quarter. Both EWR and JFK pick up many of the census block groups assigned to LGA, but it is above all JFK that gains: it even regularly appears to be the best choice in some of the block groups in The Bronx even though that choice is associated with extra driving time as LGA is the more proximate option. JFK's strong showing for fare-aware passengers in this final quarter of the year is also visible in its capturing larger parts of Manhattan and a number of block set groups west of the Verrazano-Narrows bridge, otherwise firmly in the catchment area of EWR.

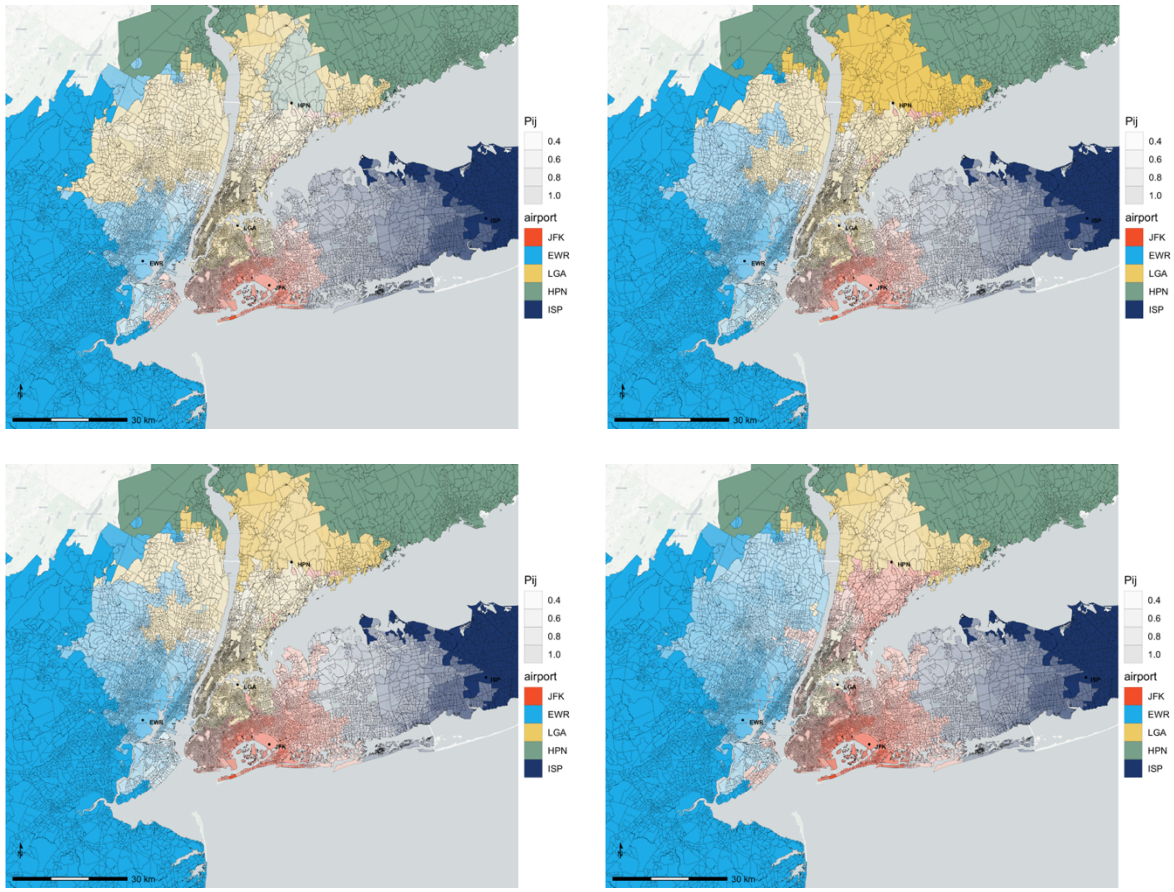


Figure 23 - Catchment areas associated with the fare utility choice set during the midday time window on Sundays for the four different quarters of the year (Q1 – top left, Q2 – top right, Q3 – bottom left, Q4 – bottom left).

4.5 Conclusions

The major objective of this paper has been to report on the development of an analytical framework that allows exploring spatio-temporal variability in MAR airport catchment areas. An ancillary objective has been to show how airport choice, and therefore catchment areas, may also depend on passengers' preferences for a specific utility set. The framework was applied to the case of domestic air travel for passengers departing from the New York metropolitan area, but we stress that results shown in this paper are above all illustrative. Even though we have drawn on the survey-based MAR literature to enhance the model's ground truth, (1) our modelling approach is rooted in assumed behaviour and therefore needs substantiation, while (2) our approach to of airport accessibility and the airport utility sets can clearly be extended and refined. For example, each of the variables (choice sets) was given an equal weighting in the choice sets (overall choice set), while the effect of distance was assumed to be linear with a maximum of 60 minutes. Moreover, the focus on domestic travel implies that LGA assumes a more central role than might be intuitively the case given the major international component of the New York metropolitan area in general and its air transport market in particular. Furthermore, accessibility could be more comprehensively measured from a multi-modal and/or cost-based perspective (e.g. Ameen & Kamga, 2013), air fare could be extended with other elements such as the price of baggage-cart rental, parking fees, and even what you pay for a coffee at a major chain such as Starbucks, etc. (https://thepointsguy.com/guide/most-and-least-expensive-us-airports/), while additional dimensions of MAR airport choice could have been added.

Many other examples of possible alterations could therefore be listed, but the key point is that the implementation of the model can clearly be improved and refined on numerous fronts. Nonetheless, we believe that some of the patterns serve to show that spatio-temporal variability matters and that our model – either in its present or in a revised form – can help capturing these. Results can then be formatively interpreted in light of the literature on airline and airport choice. For example, the results clearly corroborate the importance of complex patterns of traffic congestion as well as seasonal variations in airports' fare structure. Addressing some of the operational limitations of the model as put forward in this paper constitutes a first obvious avenue for further research: analysing which measures of accessibility and utility are most pertinent, and to what degree these matters. This could be based on previous survey-based research and would enhance the ground truth of the patterns shown. However, there are also options for further research that build on the model (either in its present or a refined form).

One of the advantages of this modelling approach is that it can relatively quickly be deployed across empirical settings, which contrasts with survey-based approaches. In the course of this research we made similar calculations – without much additional research time or effort – for the Los Angeles and San Francisco MARs, and we could also fairly easily do this for other US settings as well. Given the right data, we could add the international component to the connectivity characteristics choice set which would allow for a more realistic analysis of the New York MAR as, or even develop a global analysis of MARs. Based on this, our approach could be used as the starting point for comparative research into the operation, outline, and dynamics of MARs and thus help addressing Fuellhart and O'Connor's (2019) observation that to date very little research has taken a comprehensive global view of the assortment of component airports in MARs. At the same time, such research would need to consider that MARs are contextual. For example, in the case of Tokyo, where intra-urban travel is comparatively more difficult than in some cities in the United States, the MAR de facto includes airports that can be reached within two to three hours.

Furthermore, there are idiosyncratic situations when it comes to distance, as in the case of São Paulo, which until 2012 had its international airport located about 100km from the city centre. And finally, in follow-up research we would like to develop a web-based tool that allows passengers, based on their location, time of travel, and possibly their destination to choose for the 'optimal' airport. This tool could also easily allow passengers to make choices, e.g. by assigning intuitive weights to the different utility sets and/or choosing for their prefer mode of access. Such a tool could, in turn, also be used by airports to assess the potential effects of different scenarios of change in whatever process drives airport choice.

CHAPTER 5

Visualizing the potential for transit-oriented development: Insights from an open and interactive planning support tool in Flanders, Belgium

MARs paper results in a US-wide analysis of MARs. We calculate the driving distance from each block-group in the US to every airport reachable within 2h30m drive.

Started working on StationsRadar paper with my colleague Freke Caset.

5.1 Introduction

Cities and regions around the world are pursuing a variety of policy and planning strategies in order to curb the adverse impacts of car-centric urban systems. One of these strategies is 'transit-oriented development' (TOD). This planning paradigm pursues a purposeful concentration of urban development around transit stations in order to support transit use and other environmentally more sustainable travel modes such as walking and cycling (Ibraeva et al., 2020). In Flanders (the northern and Dutch-speaking part of Belgium), the spatial development principles of the TOD paradigm are firmly embedded in current policy and planning debates (Boussauw et al., 2018). This is informed by environmental and socio-economic sustainability goals, such as transitioning to a more sustainable mobility system and safeguarding the accessibility of the region's major urban-economic centres.

Against this backdrop, the research presented in this paper reports on the development of a novel, open and interactive planning support tool named 'StationsRadar'. The tool classifies as an 'accessibility instrument' (te Brommelstroet et al., 2014; Papa et al., 2016; Cecilia Silva et al., 2019) as it is intended to support integrated land use and transport strategy-making at railway station locations. We developed the tool in close dialogue with Flemish policy and planning stakeholders by drawing on the experiential case study research strategy that was recently proposed for planning research by Straatemeier et al. (2010) in this journal (see also Straatemeier, 2019 and many of the contributions discussed in Silva et al., 2019). By invoking this methodological approach, we subscribe to the widely shared contention within current debates on planning support systems (PSSs), and on accessibility instruments in particular, that instead of developing ever more technically advanced tools, more research is needed that probes actual user experiences and expectations and explicitly involves the local planning and political-institutional context in the development process (Balducci & Bertolini, 2007; Cecília Silva et al., 2017; Cecília Silva & Larsson, 2018).

Besides the practical pursuit of providing the Flemish regional planning practice with an empirical tool to better inform current TOD planning debates, the work presented in this paper has a clear-cut methodological objective: we aim to contribute to a better understanding of how to develop and design accessibility instruments for TOD planning purposes. We particularly focus on a branch of TOD planning support tools that has derived from the literature on 'node-place modelling' (originally Bertolini, 1999) (some examples include Balz & Schrijnen, 2009; Singh et al., 2017; Caset et al., 2018, 2019; Groenendijk et al., 2018; Papa et al., 2018; Vale et al., 2018; Nigro et al., 2019). These empirical station assessment tools are intended to support TOD planning processes by visualizing the performance of station locations on a range of transport ('node') and land use ('place') accessibility indicators. However, while the vast majority of these studies foreground, or at least hint towards, the relevance of their developed tools for planning practice, to date surprisingly little work has been undertaken to verify these claims.

By this token, this paper aims to contribute to a better understanding of the added value of this type of TOD planning support tools for planning practice, and this by deriving insights and recommendations from our experiential approach applied to the case of StationsRadar. We particularly focus on aspects of tool 'usability', i.e. the perceived ease of use and performance of the tool functionalities such as user friendliness, data quality and visualization, transparency and communicative value (Pelzer, 2017). In the context of this research we also examined tool 'utility', i.e. the 'fit' of the tool with the phase of the planning process and the scale of the planning issue (Ibid.). However, in order to keep this paper self-

standing, we mainly focus on the outcomes of our usability appraisal in what follows. We refer to authors (2019) for a focused discussion on tool utility in the context of Flemish regional planning.
The remainder of this chapter is structured as follows. In Section 5.2 we provide more background on the type of TOD tools that StationsRadar builds on. We also elaborate on the rationale and motivation behind invoking an experiential case study research strategy in light of this research. Section 5.3 discusses the methods: we introduce the reader to the StationsRadar beta version and clarify the experiential approach and workshop protocol and set-up. In Section 5.4 we elaborate on our main findings and we clarify the technological development trajectory of the tool. We wrap up this paper with a discussion and a conclusion in Section 5.4.2, and formulate specific recommendations and challenges for future research efforts along these lines.

5.2 Background

5.2.1. Visualizing the potential for TOD: An overview of empirical station assessment tools

The StationsRadar tool builds on the 'node-place modelling' literature. Bertolini (1999) introduced the model as "an analytical tool to help identify the potential for public transport-oriented urban-regional development", and applied it to the Amsterdam and Utrecht urban agglomerations. In its most basic guise, it takes the shape of a simple x ('place') and y ('node') diagram, in which different indicators are translated into a node and place index by means of multi-criteria analysis. The node index is operationalized as the transport accessibility of a railway station, while the place index is conceived as a cumulative accessibility measure capturing the intensity and diversity of activities in the 'station area' (usually defined as a station's walkable precinct). The place index is typically interpreted in terms of the 'D' ingredients of TOD planning – 'density', 'diversity' and 'design' – as first proposed by Cervero and Kockelman (1997) and later extended by Ewing and Cervero (2010).

Over the past two decades, this literature has produced a proliferation of academic and non-academic studies in which empirical station assessment models have been developed. Typically, these studies develop visual renderings of 'node' and 'place' performance levels, taking the shape of polar graphs in which relative performance levels are plotted on scaled axes with a common origin. Figure 24 provides a non-exhaustive overview of these type of visual renderings. Some recent examples, all of which were developed in The Netherlands, are the 'kite model', the 'node-place diagram', and the 'butterfly model'. The former comprises five dimensions. Alongside some typical node- and place- like features (such as 'position of the station in the public transport network', 'multimodality' and 'urbanization of station area'), additional dimensions were added such as the presence of services at the station. The 'node-place diagram', in turn, divides the standard cartesian diagram into four axes. Finally, the 'butterfly model' represents a visual rendering with six axes, reminiscent of the wings of a butterfly. The left 'wing' includes all node-related dimensions and the right wing place-related dimensions.

In addition to these applications, the 'node-place-experience' model (Groenendijk et al., 2018) adds indicators reflecting the traveler's experience at the station (in terms of comfort, ambient elements, and personnel presence). Meanwhile, Vale et al. (2018) extended the model with a TOD 'design' dimension, reflecting the 'walkability' of the station areas. The web diagram introduced by Singh et al. (2017) also quantifies the walkability and 'bikeability' of the station area, alongside dimensions such as 'user-friendliness' and 'passenger load' of the transit system. Two other recent examples include the work of Papa et al. (2018) and the triangular polar graph introduced by Nigro et al. (2019). Similar to the framework developed by Caset et al. (2018), the latter also visualizes the impact on the 'place' dimensions for different station area sizes.

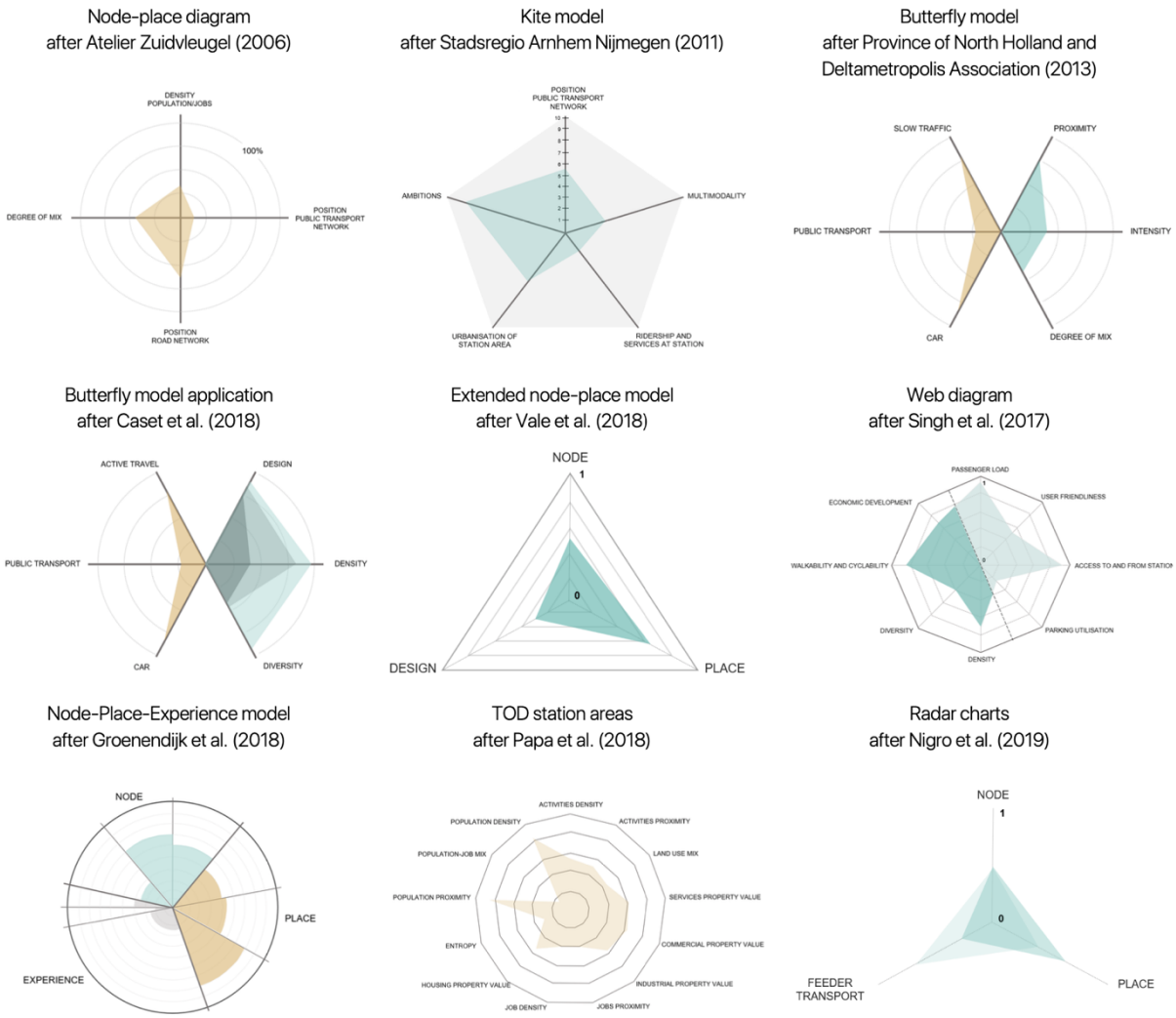


Figure 24: Non-exhaustive overview of polar graph visualizations in the TOD literature

The key assumption underpinning these TOD support tools is that the development potential of railway station locations can be derived from the empirical evidence provided by these transport and land use (and sometimes additional) indicators. The underlying assumption, then, is that these visual renderings will help communicate these findings to policy and planning professionals in order to shape strategic TOD planning and policy discussions. Surprisingly, however, this key assumption is rarely validated in close dialogue with the intended users of the applications. Two notable exceptions are Duffhues et al. (2014) and Kickert et al. (2014). Both papers report on a serious game named[69] ''SPRINTCITY' that is built around an intervention model drawing on node-place modelling principles and indicators. The game was developed through a continuous feedback loop between its players and the game developers. The nature of this application is nonetheless different compared to the applications discussed in this paper, as SPRINTCITY is a predictive (instead of a descriptive) model.

[69] Also known as applied game, it refers to a game designed for a primary purpose rather than entertainment.

5.3 What works, and why? Accessibility instruments and experiential workshops

Few planning support instruments commonly discussed in the literature (node-place models included) are explicitly validated in close dialogue with their intended users (Balducci & Bertolini, 2007; Pelzer, 2017; Straatemeier, 2019). This in turn reveals a lack of cross-fertilization between the output of applied academic research and actual planning instruments which hampers the integration of scientific and practical knowledge (Balducci & Bertolini, 2007).

For the particular case of 'accessibility instruments', this contention has been voiced frequently over the past years (Cecília Silva et al., 2017; Cecília Silva & Larsson, 2018; Cecilia Silva et al., 2019). As Papa et al. (2016) explain, accessibility instruments are "a type of planning support systems (PSS) designed to support integrated land-use transport analysis and planning through providing explicit knowledge on the accessibility of land uses by different modes of transport at various geographical scales". While there is an extensive body of work on the development and classification of accessibility measures, usefulness assessments of these methodological advances as perceived by their intended users (planning and policy professionals) remain thin on the ground (Silva et al., 2017). As a result, a plethora of accessibility instruments are produced, often based on abstract ideas that are far removed from actual practice and that lack a clear, shared understanding of the needs and demands of the specific planning context at hand (Ibid.).

In line with the studies mentioned above (see Cecília Silva et al., 2017; Cecília Silva & Larsson, 2018; Cecilia Silva et al., 2019), we argue that in order to address this type of research questions ('What works?' and 'Why does it work?'), academics need to engage with practice and submit their findings to explicit testing in close cooperation with relevant stakeholders. Recent efforts in this direction were put forward by Straatemeier et al. (2010), and have since been applied in different research settings and geographical contexts (see te Brommelstroet et al., 2014; Silva et al., 2019). The methodology put forward by Straatemeier et al. (2010) is coined 'experiential case-study analysis' and draws on theories and methods of 'experiential learning' as articulated in the field of education by Kolb and Fry (1974). As explained by Straatemeier (2019 p.55), central to this approach is the notion that experiential learning unfolds through "an iterative sequence of interlinked activities, with a continuous shift between reflection and action, the one nurturing the other". By the same token, an experiential research design should allow for connections between the following interlinked sets of activities in a direct and systematic way: 'observation and reflection' (O&R), 'forming of abstract concepts' (FAC), 'testing in new situations' (TNS) and 'concrete experience' (CE). In more specific terms, such a research design spiral requires a series of 'close-to-real-life' cases that allow lessons from the first case to be included in the second case and so on. In the process, researchers build on concrete experience provided by the planning professionals and aim to gradually enhance the relevance of their theoretical improvements for planning practice, and this in order to deduce meaningful insights about the underlying mechanisms that determine why particular planning innovations do or do not work.

5.4 Methods

5.4.1. StationsRadar: the beta version

The StationsRadar tool is rooted in earlier work (Caset et al., 2019) in which a 'node-place-people' model was developed and applied to all 287 railway station locations in the Flemish and Brussels railway network. Similar to the examples discussed in Figure 24, the indicators were visualized by means of a polar graph. These include railway network centrality indicators and feeder mode (bus, tram, metro, car and bike) accessibility indicators, as well as contour measures quantifying land use characteristics of the station area (densities of jobs, inhabitants and amenities, the morphological and functional mix of land use and the walkability of the built environment). Besides these 'node' and 'place' characteristics, a 'people' dimension reflects rail user-based data that was provided by Belgian National Railway Company NMBS. These data provide insight into the size of a station's catchment area, ridership numbers for different weekdays, and the profile of station users.

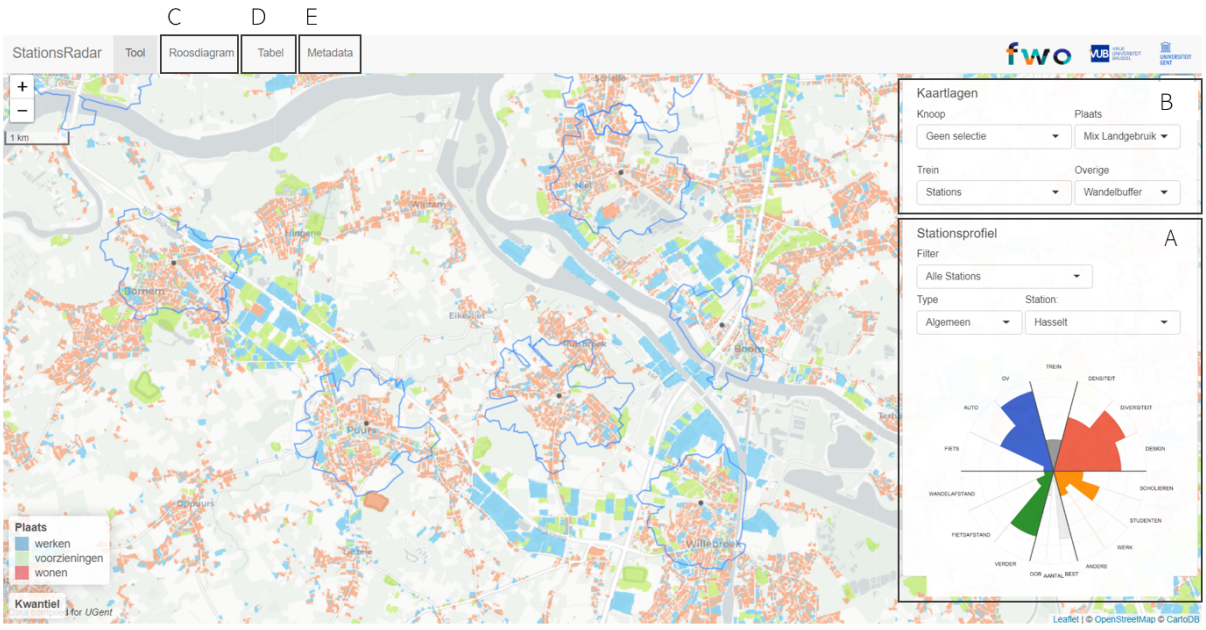


Figure 25: The StationsRadar beta version - tool components

Figure 25 shows the landing page of the beta tool version. Box A illustrates a polar graph example for the station of Hasselt. This functionality allowed to visualize one polar graph for a station of choice. Importantly, in line with the tools discussed in Figure 24, these graphs were static, in that they displayed fixed and standardized indicator scores (reflecting the relative performance of a particular station compared to all others on a scale between 0 and 10). Box B allowed the user to plot different thematic maps, while boxes C, D and E provided information about indicator calculations, a data table with the absolute indicator scores, and metadata.

The tool was developed in R, through the RStudio[70] Integrated development environment (IDE). The following R packages were used: tidyverse[71] (data collection, curation and analysis), ggplot2[72] (polar graphs), leaflet[73] (maps), and Shiny[74] (R translation to a JavaScript based interactive web app).

5.4.2. Three experiential workshops

In order to probe for the perceived usability of our tool, we devised three close-to-real-life 'experiential workshops' (Silva et al., 2019) with Flemish planning and policy professionals. According to Billger et al.'s (2017) typology of usability studies, our study classifies as a 'prototype study in a simulated setting'. The guiding question throughout the experiential process was the following: How usable is the StationsRadar tool and (how) can its usability be improved?

Three half-day workshops were organized in the planning context of three different 'transport regions': Ghent, Aalst and Leuven (see Figure 26a). These recently established regional partnerships have been devised to stimulate cooperation between different stakeholders (municipalities, public transport operators, the Flemish Government and others) on the organization and coordination of public transport networks in the region, and this in dialogue with the domain of spatial planning. In these partnerships, the strategic development principles of the TOD paradigm are expected to be translated into practice.

For each workshop, multiple station cases were selected in close dialogue with the local co-organizers. This selection was made on the basis of several arguments. First, station-specific elements played a role. We aimed for cases (i) that were the subject of current and relevant transport and/or urban planning questions, and (ii) that together formed a balanced mix in terms of their regional importance. Second, certain cases were selected based on stakeholder-specific arguments; some municipal stakeholders were deemed more experienced and 'passionate' about the topic, which could have ramifications in terms of the success of the workshop and the overall group dynamics. In terms of sampling strategy, the workshop stakeholder composition closely mimicked that of the administrative leg of the transport region council. The test users are thus representatives from: the municipalities in which the station cases are located, the Flemish Government, the Provincial Government, intercommunal organizations and public transport companies NMBS and De Lijn (the Flemish bus and tram company). Given the tool's integration of transport and land use indicators, we aimed for a balanced workshop presence of participants with a background in transport and spatial planning. In total, 45 participants attended the workshops.

Our workshop protocol draws on the work of te Brömmelstroet et al. (2014) and was modified in line with ideas raised by the local co-organizers. Each workshop consisted of five distinct parts:

- (A) Introduction (15').
- (B) Intuitive exercise (30'): A round-the-table exercise in which the municipal representatives were invited to introduce their station and describe its accessibility in an intuitive way.

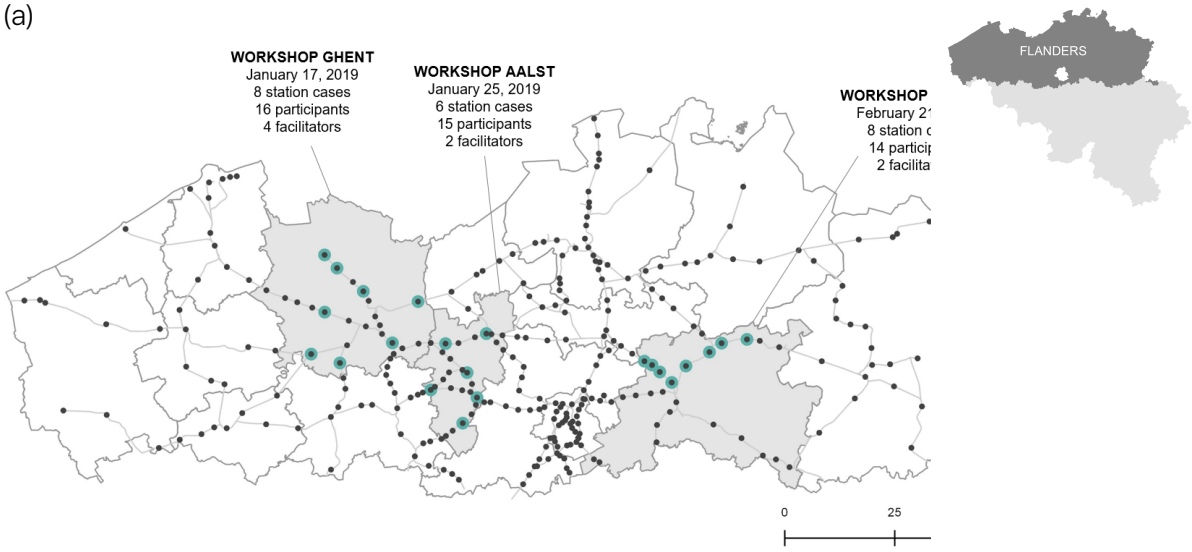
[70] Integrated development environment for R (https://rstudio.com).
[71] Collection of R packages developed by Hadley Wickham's team, that follow a design philosophy and grammar of data and graphics (https://www.tidyverse.org).
[72] Part of the tidyverse package (see above).
[73] Open source JavaScript library that allows the creation of interactive maps (https://leafletjs.com).
[74] R package that allows creating interactive web apps directly from R (https://shiny.rstudio.com).

- **(C) A hint of theory (’45):** A clarification of TOD concepts, the method of node-place modelling and the StationsRadar tool functionalities.
- **(D) Tool testing (135’):** Participants were assigned to worktables (see Figure 26b)) that functioned as focus groups, and that hosted a balanced composition of participants in terms of organisation, background and expertise. Each focus group discussion centred on topical planning questions that pertained to the TOD potential of each station (area) case. In order to address these questions, participants had to consult both the tool and each other’s perspectives. Each focus group was moderated by a facilitator of our team who actively steered the discussion to zoom in on relevant usability statements and hypotheses.
- **(E) Survey (’15):** A post-workshop survey.

Data was collected in the B, D and E parts of each workshop. For practical reasons we will only discuss D and E. The focus groups were audio recorded and transcribed verbatim. The survey contained Likert-scale statements rated 1 to 5 (from ‘strongly disagree’ to ‘strongly agree’). In total, 43 surveys were completed. The survey design drew on the work of Champlin et al. (2019) in that it focused on the following dimensions: the participants and their background, the perceived quality of the workshop process at the individual and group level (evaluating general satisfaction, insight, communication, shared language, consensus-building and efficiency gains), tool usability (evaluating transparency, credibility, output clarity, focus, level of detail, etc.), and tool utility (evaluating the added value of StationsRadar in the Flemish planning context of the transport region). On a total of 40 statements, the survey included 22 usability statements, 14 of which specifically focused on the polar graph data visualizations, and 6 on the overall tool functionality.

In line with an experiential research design, usability hypotheses were continuously revisited as an input for each successive workshop. In other words, hypotheses that were raised during the first (or second) workshop were (re)introduced by the facilitators during the second (or third) workshop. Importantly, in contrast to the work of Straatemeier (2019), the tool was not modified in between workshops, as we lacked the resources to do so within the short timespan in which the workshops were planned. This has an important methodological drawback in that the received feedback is not grounded in actual before and after experimentation. However, as the protocol was uniform across the workshops, our approach allowed us to aggregate our findings and formulate robust usability expectations. Moreover, as will be illustrated, given that the feedback across workshops was consensual, our findings can be interpreted straightforwardly.

(a)



(b)

Figure 26: (a) Geography of the workshop cases in the Flemish and Brussels railway network, (b) Illustration of a worktable setting

5.5 Findings

In this section, we summarize the most important usability outcomes that we derived from the workshops (Section 5.4.1), followed by a clarification of how we transformed the beta version of the tool accordingly (Section 5.4.2).

5.5.1. Usability insights

a) The experiential learning process

We start with a chronological account of the insights collected throughout the entire experiential process, from our perspective as academics. We draw on the focus groups to illustrate particular findings by means of citations and discuss the most relevant survey findings. We refer to Table 10 for a schematic visualization of the most important feedback in terms of tool usability and the workshop process.

The experiential learning process started with a number of observations and reflections (O&R) that took shape during the preparatory meetings leading to the first workshop. Our local co-organizers expressed a strong interest in the tool for a number of reasons. First, an academic and alleged 'politically neutral' setting in which a sample of crucial regional stakeholders would be joined under the banner of TOD, was deemed highly interesting as it would allow our co-organizers to probe for the stance of the participants with respect to this new policy principle. Also, whereas our co-organisers hypothesised that the tool might introduce a 'common ground' to support supralocal discussions about station development potential, concern was also raised that the indicators were 'very mathematical'. After two tool stress-tests with our university colleagues and a further refinement of the tool and the polar graphs (FAC), StationsRadar was ready to be tested during the first workshop in Ghent (TNS).

At the time of the workshop, the transport region of Ghent had just been established. The concrete experience (CE) on the basis of which the tool was validated was therefore largely stakeholder-specific, instead of it being a cohesive planning practice with well-defined roles. Nonetheless, some clear observations and reflections (O&R) in terms of tool usability could be made. First, the tool was deemed most relevant for the 'supralocal stakeholders' (the mobility providers, the intercommunal organizations and the Flemish and Provincial Governments). A variety of uses on the regional scale were envisioned: to 'better inform regional allocation decisions', 'help developing a hierarchy of nodes', 'help integrating the different layers and modes of public transport in the region' and 'function as a communication tool between stakeholders'. However, the added value of the tool at the local, municipal, level seemed less evident. While many participants stressed the need for empirical evidence as an input for local strategy-making, the polar graphs were deemed insufficient at this stage, mainly in terms of level of detail. Most municipal stakeholders stated that the absolute figures provided in the data table were (far) more relevant than the relative scores displayed in the graphs. A second reflection concerned the lack of interactivity of the tool and, more specifically, the observation that users could not plot polar graphs as a function of their own desired station selections: 'It would make more sense if we could compare stations of a similar size and order'. Additionally, the tool should allow plotting multiple graphs next to each other, fostering the ease of visual comparison. Third, we observed and experienced how the NMBS rail user-based data revealed novel and meaningful insights. This was especially the case for representatives of smaller municipalities who generally lack the resources to frequently update mobility plans and organize passenger counts or conduct user-based surveys.

With the above reflections in mind (FAC), we embarked on the second workshop in Aalst (TNS). In contrast to the previous case, the transport region of Aalst was established in 2016 as a pilot project. The concrete experience (CE) of the workshop participants was therefore more developed in terms of there being a collective practice. In general, most participants found the idea behind the tool very strong, referring to the integrated approach of mobility and spatial planning and to the 'stimulus' it could give to 'thinking more regionally'. Similar to the previous workshop, the difference in perceived usability between the local and the regional governance scales was quickly raised. However, during one of the focus groups a discussion arose about how the tool's usability could be improved for local stakeholders, and how this in turn could benefit the transport region's functioning. As a mobility expert explained: 'If the tool would allow for flexible polar graph comparisons between municipalities, then it might foster inter-municipal dialogues in which certain measures taken and their actual impact are compared and discussed. For example, if a municipality introduced toll parking at the station, it would be interesting to see, also for neighbouring municipalities, how this affects particular parts of the graph. In this way, the tool could foster a bottom-up and peer-review dynamic that could reinforce the transport region'. Additionally, a series of interesting improvements in terms of polar graph visualization were proposed, such as displaying the absolute data when hovering over a slice of the graph. Or, as one spatial planner proposed: 'It would be great if we could make selections of stations based on one particular theme, such as 'ridership'. In that way, you could easily select stations with similar ridership characteristics, plot their graphs and examine how and why they are performing differently'. These usability statements reveal a similar need for interactivity as was expressed during the first workshop. Another point that was also raised earlier concerns the difference in expertise and resources between smaller and larger municipalities. As stated by an Alderwoman responsible for Mobility and Public Works: 'The problem is that, and I mainly speak on behalf of the rural municipalities, whenever you have all that information, you need to be able to work with it. You need to have the manpower to get started with it and draw conclusions from it'. This statement resonates with the perceived complexity of the beta tool version by many participants. As one mobility expert put it: 'After today's workshop it became clear to me how the tool comes close to the complexity level of our transport models or ArcGIS. In other words, you will always need an operator'.

We concluded that the second workshop led to some innovative usability suggestions, and that our observations and reflections (O&R) were largely in line with those of the first workshop. We also experienced how some stakeholders (such as the bus and tram operator and some municipalities) offered to contribute to the tool by providing additional data, which inspired us to reflect on a tool design that could cater for increased user involvement in this direction.

With the above usability hypotheses in mind (FAC), we embarked on the final workshop in Leuven (TNS). Although this transport region had just been established, a large share of participants was experienced in working together on this regional scale (CE) due to their involvement in another regional project. Similar to the previous workshops, an important observation was that participants requested more flexible station comparisons, and that they stressed the importance of the absolute numbers over the relative graph scores. For example, a municipal mobility expert asked: 'But why did you opt to compare stations with each other? This diagram totally contrasts with how we are used to look at things. You look completely different at those numbers. We always start by looking at the absolute numbers, the inflow: how much and how do people get there etc. But these diagrams... It's all so relative'. Along with the above, suggestions were made to alter the way in which the relative scores

were normalized, and ideas for additional indicators were proposed such as a 'design for all' indicator (reflecting the accessibility of the station and bus stops for disabled persons), an indicator reflecting perceived safety of the station area and one reflecting the level of road congestion in the vicinity of the station. However, other participants questioned the need to further expand the amount of information included and would rather distil the most relevant indicators only. A final observation in line with the previous workshops was the strong interest for the NMBS user-based data. For example, a Provincial policy officer responsible for spatial planning reflected 'how great it would be if the data about the catchment area sizes could also be visualised spatially, let's say by using raster images so there is no privacy problem'.

b) Survey findings

In terms of the graphs, and in line with the above findings, the majority of participants stated that they valued this type of polar graph visualization, provided that some of the limitations (such as lack of interactivity and the importance of the absolute data figures) would be tackled. Similarly, when asked if the polar graphs are 'too abstract', most people stated that this is not the case, 'as long as you fully realize what you are comparing and what the scores really mean'. Or, as one participant noted: 'For me it's all about the scale of abstraction. It's fine to compare between stations at the regional scale, but on the level of let's say one station, a polar graph is removed too far from reality and in this case, I am more in favour of the combination of multiple tools to approach reality'. And also: 'In order to make sense of this complex matter, I don't think you can proceed differently than through an abstraction of reality'. For another statement that probes for aspects of indicator operationalization, we received a high number of blank responses, revealing that our workshop set-up did not provide enough time for most attendees to respond in a well-informed way. As one participant noted: 'We should be able to work with the tool for a longer period, let's say a week, in order to give more grounded feedback'. Suggestions for extra dimensions and indicators were nonetheless made, and are mostly in line with the ones raised during the focus groups. When querying the communicative strength of the graphs, opinions were divided. Those who do not agree mostly refer to the extensive knowledge that is required to interpret the indicators correctly, and therefore argue that StationsRadar is 'definitely not a quick visualization tool'. In a similar vein, some state that the communicative value is only tangible for 'professionals'.

In terms of the general assessment of tool functionality, the following observations can be made. First, the majority finds the tool user-friendly and does not think important cartographic material is missing. As for the latter, some interesting suggestions were nonetheless made, such as a layer visualizing the expected demographic change in the region, and a layer that informs the demographics of the inhabitants of the station area (age, income, ...). Second, opinions were divided in terms of the perceived transparency of the tool in terms of data and indicator operationalization. And third, the critiques mentioned above in terms of tool interactivity were also very evident from the survey responses.

	+ (deemed positive)	– (deemed negative)
Usability: Polar graphs	The overall principle of visualizing these empirical data by means of polar graphs The integration of data pertaining to the domains of mobility and spatial planning The user-based perspective on railway accessibility (besides the conventional 'node' and 'place' dimensions)	The relevance of these graphs for stakeholders operating at the 'local' (i.e. municipal) scale The high level of prior knowledge needed to interpret the graphs correctly The absence of some important indicators (e.g. station accessibility for disabled people and perceived safety) The lack of tool interactivity (users want to visualize and compare the graphs for tailored sets of stations, and users want to visualize the absolute indicator scores when hovering over a graph) The lack of additional data visualization possibilities (users want to visualize and compare stations of similar size and order (e.g. stations with similar ridership performance), and users want to visualize multiple graphs at the same time) The normalization method used to calculate the relative scores
Usability: Overall tool	The user-friendliness of the interface The tool transparency (e.g. the representation of the absolute numbers in data tables)	The absence of some spatial data layers (e.g. the expected demographic growth and the socio-economic composition of households in the station areas)
Workshop process	The inter- and transdisciplinary workshop set-up The establishment of a shared professional language The establishment of constructive social dynamics The establishment of a better understanding of the viewpoints of some other stakeholders	The lack of time to fully grasp and discuss some of the indicators and their operationalization

Table 10 - Summary of the overall feedback in terms of usability and the workshop process

5.5.2. A renovated StationsRadar tool

Drawing on these usability insights, we thoroughly renovated the tool. Figure 27 illustrates the renovated tool's different components, which we now briefly discuss.

The majority of participants expressed a desire to plot multiple polar graphs simultaneously, and to plot the scores for tailored and flexible sets of stations. In order to live up to these expectations, we had to rethink the way in which the polar graphs were created. While ggplot2 offers much interesting features, it does not allow for this kind of flexibility. We therefore opted for an open-source Javascript[75] framework by using 'Vue.js[76]', 'Vuetify[77]', and the JavaScript libraries 'D3.js[78]' and 'Highcharts[79]'. The polar graphs were designed using D3.js and Highcharts. The relative scores are now calculated reactively, so that the performance values are scaled relative to the particular group of stations selected. Also, when hovering over the diagram the absolute performance values are shown along with an indicator description. Figure 27d provides an illustration for the 27 stations that are located in the transport region of Aalst.

Besides this intervention, we also tackled the user feedback pertaining to the ability to plot selections of stations for one particular theme (i.e., one particular slice of the polar graph). To this end, we invoked line chart visualizations that display the absolute indicator scores in a more informative way (see Figure 27c). These charts give a quick overview of indicator performance and distribution across the stations selected. Figure 27c illustrates the performance of a group of 12 stations that are located along a rail corridor, and this for the theme of 'rail-based accessibility' which consists of six indicators. When hovering over the plot, the absolute values are shown. These charts were developed by drawing on the 'spline with inverted axes' template in Highcharts and D3.js.

Besides these charts, the absolute data can also be consulted in the data tables as illustrated in Figure 27e. These tables are reactive in that the user can easily search and group records. In line with user recommendations, we also incorporated additional indicators where feasible. For example, recent data of station car and bike parking utilization rates were provided by one of the organizations and were added to the table.

The 'maps' tab serves to visualize geographic datasets in order to enhance the interpretation of the different data graphs. Figure 27a and Figure 27b provide some illustrations. The former map displays a vector layer classifying all Belgian stations according to their 'transfer centrality' (authors 2019) in the railway network, whereas the latter displays a raster map showing the density of 'regional amenities'. All maps are zoomable and additional data attributes can be visualized when hoovering over the map.

[75] High-level, multi-paradigm programming language (https://www.javascript.com).
[76] Open-source model–view–viewmodel JavaScript framework for building user interfaces and single-page applications (https://vuejs.org).
[77] Material Design component framework for Vue.js (https://vuetifyjs.com).
[78] JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of Scalable Vector Graphics, HTML5, and Cascading Style Sheets standards (https://d3js.org).
[79] Interactive and scalable JavaScript based graphs (https://www.highcharts.com).

Figure 27: The components of the renovated StationsRadar tool

5.6 Discussion and conclusions

This paper reported on an experiential approach to the development of a TOD planning support tool in Flanders. At the root of this project was the observation that few of the accessibility instruments commonly discussed in the literature (node-place modelling applications included) are explicitly validated in close dialogue with their intended users. This is surprising, as the majority of node-place based studies touch upon the interface between planning research and practice and foreground, or at least hint towards, the usefulness of their empirical outcomes to (a variety of) stakeholders involved in TOD planning.

In order to help bridge this gap, we extended the work of Straatemeier (2019) and Silva et al. (2019) by organizing a number of experiential workshops in which the recently developed StationsRadar accessibility instrument was tested and subsequently revised on the basis of the concrete experience of policy and planning stakeholders actively involved in the Flemish transport regions. The development process from the beta to the renovated version that is now published online (see https://stationsradar.ugent.be), can be considered part of another loop in this experiential learning process, as we revisited and altered the abstract concepts (the graph and map visualizations) and produced a version that is now ready to be submitted to new rounds of testing, albeit in a real-life context.

The usability observations and reflections that we discussed in this paper bear direct relevance for the well-rehearsed practice of developing empirical station area assessment tools for TOD planning. While each planning context is unique, it may well be the case that our usability recommendations are, to a certain extent, transferable across cases. Below, and by way of concluding this paper, we therefore summarize the most important general usability recommendations emanating from our study. In the process, we reflect on the broader technical and methodological challenges that come with implementing these in practice.

1. Interactive and diversified data visualizations: There was a clear consensus that the data and the derived indicators needed to be visualized as interactively as possible, allowing users to draw and compare graphs on the fly for tailored sets of railway station locations. Additionally, participants expressed a need to consult the data by means of multiple, diverse visualization modes. For example, the line charts serve a different purpose compared to the polar graphs in that they quickly provide absolute numbers and data distributions for tailored sets of indicators, whereas the polar graphs provide a more generic station profile reflecting aggregated, relative, performance levels. These observations are revealing in that none of the TOD support applications discussed earlier have incorporated interactive elements, nor have they (or do they seem to have) experimented with multiple data visualization techniques beyond the traditional polar graph standard. As a corollary, we believe that future work along these lines (i.e., work that develops TOD support tools that depend on strong visual cues) may benefit from a closer engagement with the field of visual analytics (dealing with visual and interaction metaphors and semantics) (see Andrienko et al., 2010 for a fuller discussion).

2. Transparent disclosure of data and actor-mobilising momentum: We experienced that it is absolutely crucial to transparently communicate the absolute numbers behind the polar graph visualizations. While this finding may not surprise, it does provide food for thought since most of the TOD applications discussed earlier stop short of this level of transparency that seems needed to

meaningfully support TOD planning. We also experienced that the open disclosure of data from different organisations instigated other stakeholders to also contribute to the platform by disclosing their own unique data. While this actor-mobilising momentum arguably signifies one of the most valuable achievements of this research project, it also brings about substantial challenges in terms of data curation. Although we devised a standardized contact sheet allowing users to get in touch if they wish to contribute, in an ideal scenario, users would be able to modify and save data records directly in the tool, thus pushing the level of tool interactivity – and ownership – to the highest possible extent. Arguably, the easiest way to accommodate this level of interaction (i.e. add, remove and edit records) implies creating a user-based portal that is supported by a full R Shiny/Vue.js integration. Such an approach would be similar to the current set-up, with the difference that R Shiny would not only be used to visualize data, but also to collect and curate the data. This approach, in turn, generates significant challenges in terms of data quality control, data integrity and in terms of resources (a dedicated server and backend development would be needed) (see also Haklay, 2010 for a fuller discussion in light of volunteered geographical information).

3. Integrating 'hard' and 'soft' data and crowdsourcing aspirations: The previous point resonates with the desire that was voiced by many participants to visualize additional 'soft' or qualitative data (Billger et al., 2017) that would pertain to aspects such as station area safety, comfort and inclusivity. For the case of StationsRadar, these data may be gathered by means of crowdsourcing techniques. Such an intervention would expand the planning support tool with a dynamic 'sounding board' functionality, displaying crowdsourced 'soft' station accessibility data as provided by citizens. The potential of such an approach was raised recently by Bertolini (2017) who hints at the importance of including non-expert planning stakeholders in experiential learning processes, possibly by means of web-based interaction. By the same token, Silva and Larsson (2018) recently made a plea to connect the different contexts and uses of the accessibility concept (i.e. academic, policy and planning, and every-day life) in a more systematic way. Such a future research avenue in which 'traditional' datasets are integrated with crowdsourced data will arguably require a more intensive engagement of the empirical TOD planning support literature with participatory approaches to mapping and GIS and with critical discourses on 'smart cities' (Batty, 2012).

CHAPTER

6

Conclusion

"If you go back a few hundred years, what we take for granted today would seem like magic - being able to talk to people over long distances, to transmit images, flying, accessing vast amounts of data like an oracle. These are all things that would have been considered magic a few hundred years ago." - Elon Musk

In the introductory chapter of this dissertation I provided an overview of the outline, overall objectives, and the different chapters. The introduction also discussed the challenges in data-driven research and in the process from data collection, to curation, analysis, and visualization. In this final chapter I will first delineate the current state-of-the-art of my research (i.e., Chapter 2 to 5) together with a summary of the main findings. This is followed by a discussion of the limitations of the research presented in Chapter 2 to 5, after which I conclude with a discussion of potential avenues for future research.

6.1 State-of-the-art and summary of findings

In this dissertation I tackled two main topics in transport geography: (1) the development of data-driven tools to minimize data complexity and (2) the opportunities engendered in data democratization. To a certain extent it is difficult to isolate both topics as they are heavily interconnected. That is, the development of data-driven tools has the potential outcome of democratizing data. At the time of its development, SKYNET was the only R package able to handle air transport data available on the Comprehensive R Archive Network (CRAN).[80] When taking a broader perspective (i.e., transport geography at large) there are only few packages developed to handle transport data, most of them focusing on handling and parsing GTFS[81] data.

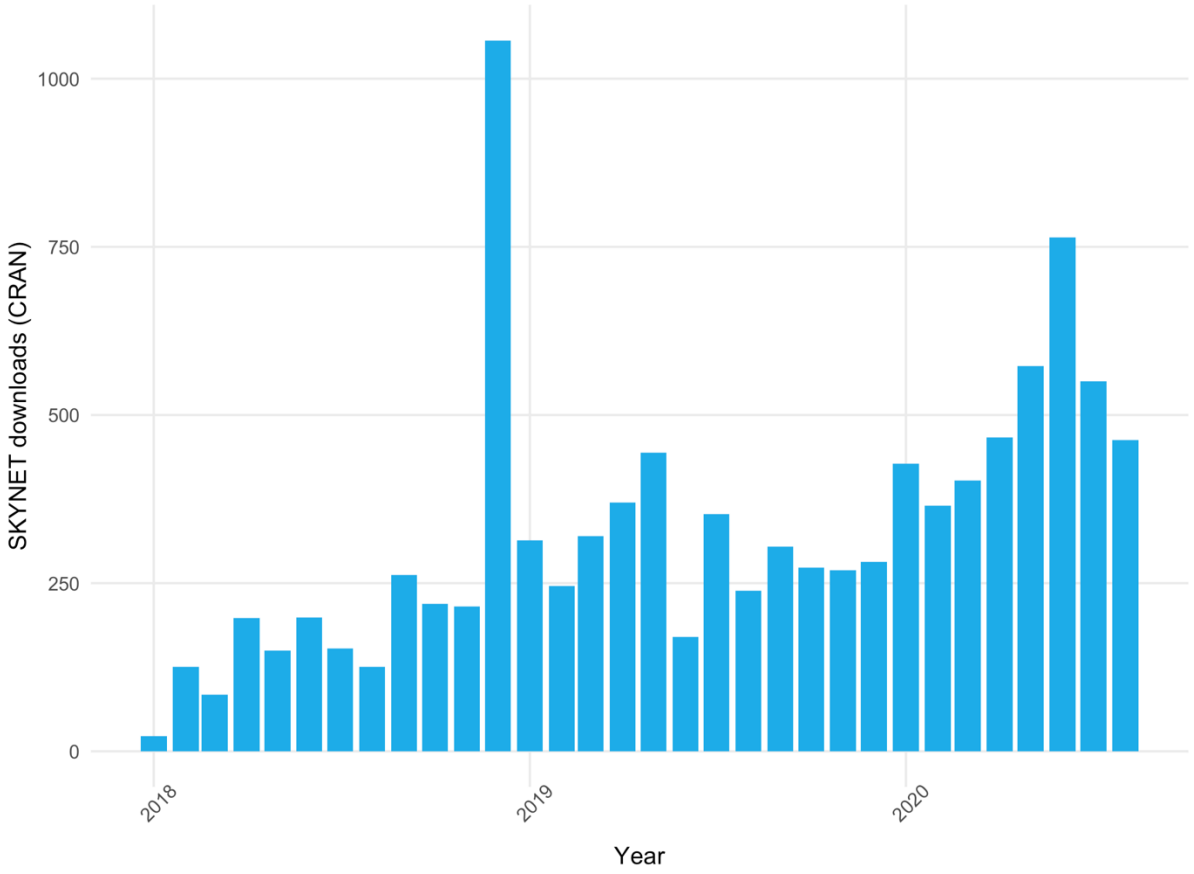


Figure 28—SKYNET downloads per month. Data extracted from CRAN and do not include github downloads.

[80] R package repository.
[81] General Transit Feed Specification (https://gtfs.org).

6.1.1. SKYNET

Since the introduction of SKYNET in 2018 and the inclusion on CRAN a few months after the paper was published, our package has been downloaded more than 10,000 times (Figure 28). This metric should be enough to show there is demand for an R package capable of interacting with the BTS DB1B and T-100 datasets. Unfortunately, aside from geographical information (e.g., downloads per country), we lack access to metrics that could shed some light on who is using the package and for which purposes. The access to more detailed metrics would allow us to better understand how SKYNET helps researchers with their analysis, and where it is lacking. Is SKYNET being used exclusively to download data, or is it being used for both downloading and analysis? Is SKYNET being integrated with other systems and approaches (e.g., machine learning, databases, other R packages)? Since R packages submitted to CRAN have to follow a strict set of guidelines[82] including mandatory recurrent updates (otherwise they are automatically removed), there are substantial efforts being put into developing a usable R package. This could well be one of the reasons for the lack of R packages directed to air transport research. At the time of writing this dissertation, CRAN listed only three other air transport-related packages. Since it was first created in 2017/18, there have been considerable updates to SKYNET. Most updates were made to improve the handling of large datasets (e.g., speed, memory). However, some new functionalities were added as well, often with data democratization in mind (e.g., the possibility to download data directly from the R environment). Despite knowing SKYNET's adoption rates, it is hard to assess if the intended goals (e.g., to become backbone of a range of easily navigable tools) have been met.

6.1.2. DB1B and T-100

SKYNET's broader potential was demonstrated in Chapter 3, in which I explored the impact of potential biases in air transport research datasets by scrutinizing the BTS DB1B and T-100 datasets. First, it facilitated downloading and data collection, and second, it facilitated analysis by formatting data in a way that allowed its immediate scrutiny. While the topic of data quality seems to have gained more relevance with the rise of AI (Jordan & Mitchell, 2015), it remains sparsely addressed in the field of transport geography (Derudder & Witlox, 2005b). A quick search in Web of Science for articles with "data" and "transport geography" as a topic (i.e., title, abstract, author keywords, KeyWords Plus), reveals rather limited levels of engagement (Figure 29).

[82] https://cran.r-project.org/web/packages/submission_checklist.html.

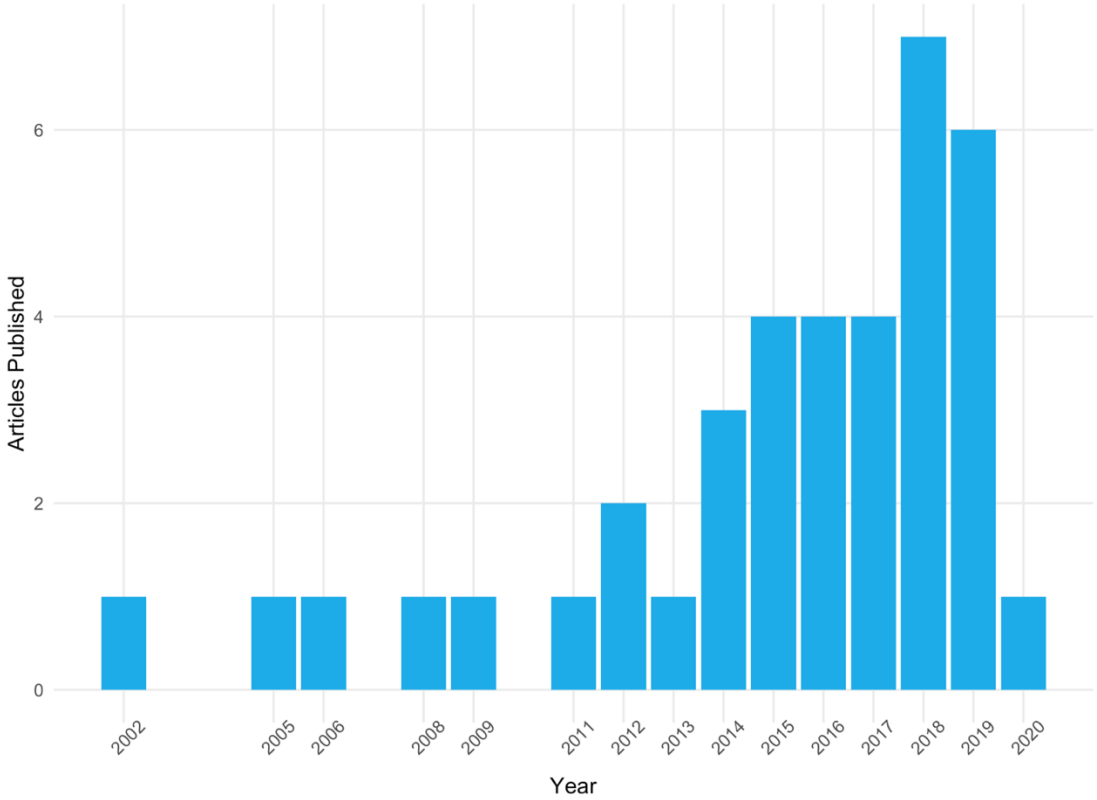


Figure 29—Research articles mentioning "data" and "transport geography" as topics. Extracted from Web of Science (01/09/20).

While I acknowledge that a proper bibliometric analysis would be needed, this brief search alone can be used as an indicator of the limited number of articles explicitly scrutinizing the nature of data in transport geography. By expanding the view to the broader field of geographic data (e.g., citizen science, social media data), a higher ratio of papers focusing on data quality can be observed. Most of these papers focus on the quality of geographic data, and most deal with social networks (Soler et al., 2012; Crampton et al., 2013; Morstatter et al., 2014; Szell et al., 2014), citizen science, and volunteered data (Haklay, 2010; Graham & Zook, 2013; Graham et al., 2013), or georeferenced Big Data (Crampton et al., 2013; Poorthuis & Zook, 2017). The interest in the quality of data and how bias can impact research may be in part fueled by the existence of tools developed to handle and analyze those data. For example, Twitter is well backed up by a solid API, alongside with packages in R (e.g., rwteet,[83] streamR,[84] tweet2r,[85] tweetR[86]) and Python libraries (e.g., tweepy,[87] twint[88]) developed for the purpose of collection and analysis. Regarding citizen science projects, OpenStreetMap is well supported by R packages (e.g., osmdata[89], OpenStreetMap[90]) and Python libraries (e.g.,

[83] https://cran.r-project.org/web/packages/rtweet/index.html
[84] https://cran.r-project.org/web/packages/streamR/index.html
[85] https://cran.r-project.org/web/packages/tweet2r/index.html
[86] https://cran.r-project.org/web/packages/twitteR/index.html
[87] https://github.com/tweepy/tweepy
[88] https://github.com/twintproject/twint
[89] https://github.com/ropensci/osmdata
[90] https://cran.r-project.org/web/packages/OpenStreetMap/index.html

OSMPythonTools,[91] osmapi[92]), as well as by solid documentation and a large user base. In the example of the BTS DB1B and T-100 datasets, and despite being repeatedly used in research, documentation is scarce and there is no API or any other form of accessing the data except through individual CSV files to be downloaded from the BTS website, or through SKYNET.

6.1.3. Spatio-temporal dynamics

In Chapter 4, attention shifted to the spatio-temporal dynamics in airport catchment areas, focusing on the case of the New York MAR. One of the first challenges we faced was the heterogeneous nature of the different data sources used, including the BTS air transport related datasets (i.e., DB1B, T-100, on-time performance). Despite sharing commonalities (e.g., they represent flights between airports), temporally they are arranged differently. For example, the DB1B is arranged per quarter, while the T-100 per month and the on-time performance dataset is grouped per hour. In Section 6.2 I will reflect on the drawbacks of using datasets with different temporal configurations, but one of the immediate issues pertains to representation. That is, if we use the DB1B (i.e., grouped per quarter) to characterize data grouped per time period (e.g., peak morning, midday), how can we ensure that the output is representative of the time period being shown (e.g., we cannot observe any air fare dynamics on a daily or weekly level, as the data is aggregated per quarter)? One of the ways to tackle this challenge was by employing a "top to bottom" approach. That is, I started with lower resolution data (i.e., quarterly DB1B) and gradually progressed to higher resolution data (i.e., hourly on-time performance). Another important aspect of this approach is that I did not try to model stated preference or airport choice, but to provide an index indicative of the "best option" depending on the different indicators (i.e., fare, connectivity, on-time performance). Regarding accessibility, I had to make concessions as to how to calculate the driving time from block groups to each airport. First, this is a computationally demanding task, and second, traffic data is often either expensive or hard to obtain. With this in mind, I decided upon creating "time bands," which would group block groups based on a driving time window to the airports being studied. One of the advantages of focusing on providing a "best option" index is that it can be relatively quickly deployed across empirical settings, which contrasts with survey-based modelling approaches. This also implies that it is possible to efficiently incorporate other variables, depending on the research question(s), in different time setups (e.g., fares per hour instead of per quarter).

At the time of writing the conclusion to this dissertation, I have made some further progress, as I was able to calculate the driving time from each block group to every airport in the US. This progress owes to being able to secure some important ArcGIS credits,[93] which in turn allowed performing the calculations, but also ensuring a setup able to perform these calculations. In this case, I had eight virtual machines[94] running for about one week, constantly querying the ArcGIS database and extracting OD matrices (i.e., driving time and distance for each given time period, from each block group centroid to each airport). The usage of eight virtual machines was due to the time costs of querying the ArcGIS database. When querying the ArcGIS database, each query can only contain five hundred origins and five hundred destinations, and it takes around three minutes to complete. If I had not used multiple

[91] https://wiki.openstreetmap.org/wiki/OSMPythonTools
[92] https://pypi.org/project/osmapi/
[93] Some ArcGIS calculations cost credits that can be purchased with every license.
[94] Virtual machines use software to emulate computer systems, by providing the functionality of a physical computer.

machines in querying the database, it would have taken about two months (i.e., not including time-outs and server downtime) to extract the entire dataset.

Stations Radar

I ended Chapter 4 by referring to the potential development of a web-based tool that allows passengers, based on a group of settings and preferences, to choose for the "optimal" airport. While this tool is still under development, during my PhD project I have worked on a similar tool (i.e., in terms of design and functionality): StationsRadar was developed to support integrated land use and transport strategy making at railway stations in the region of Flanders and Brussels in Belgium. As mentioned in Chapter 4, StationsRadar was developed in close dialog with policy and planning stakeholders, by means of workshops and by allowing the stakeholders to experiment and use the tool. The feedback gathered after several practice sessions provided the necessary tools and direction needed to develop both versions (i.e., beta, final) of the tool. As the theoretical framework involving this tool has been thoroughly discussed in Caset et al. (2019), here I will focus on the technical development and output of the tool.

In terms of simplification of complex data and data democratization, the output of StationsRadar can be divided into three dimensions: (1) interactivity—how does the tool allow users to extract, visualize, create, modify, and visualize data? (2) transparency—how does StationsRadar facilitate the access to data by increasing its accessibility? (3) usability—considering the first two dimensions, what is the practical (e.g., academic, policy and planning, and every-day life) impact of StationsRadar as a planning support tool in the context of TOD? The earlier versions of StationsRadar (i.e., exclusively written in R and Shiny) were limited in terms of interactivity. While some of these limitations were caused by the limitations of R and Shiny, we were also limited by the lack of awareness of how much interactivity was required by potential users of this tool. It might be reasonable to assume that any planning support tool benefits from interactivity, but in practice, when given too much control, users might easily be overwhelmed by both the tool and the data. The balance of how much interactivity should be given to users was an important element driving most design and programming decisions made when developing the final versions of the tool (i.e., coded in JavaScript and Vue.js). In order to find the right balance of interactivity, it was important to accommodate all levels of expertise: some users could be more familiar with planning support tools and data-driven web tools in general, while others could be more resistant to the adoption of new technologies.

In StationsRadar, ensuring access to the data's metadata was deemed essential, and access to the actual numbers behind some of the displayed graphs was pursued when possible. When aggregating and standardizing the information through the form of graphs and maps, some users expressed concerns, as the raw number, as well as the provenance of the data, were not visible. With this in mind, I added several tables that include a series of important metadata (e.g., collection, author, date), alongside raw values for the metrics displayed in graphs and maps. The usability of a tool is by nature an elusive concept as it is difficult to formally measure. Caset (2019) focuses on StationsRadar's usability and utility in the variety of uses on the regional, local, and municipal level. In this dissertation, by referring to usability, I am mostly interested in the ease of use and potential the tool has to communicate data, and therefore facilitate its interpretation. For example, when developing the radar diagrams it was important that data could be immediately interpreted and compared when selecting multiple stations (Figure 30 and Figure 31). This in turn created some challenges, namely in ensuring that: (1) there would be enough diagrams present on screen; (2) the colors would be suitable for color-blind people; (3) when hovering the cursor, more information would be displayed, showing raw values;

and (4) multiple radar diagrams would not require high computational power. These four points led me to having to tap into the source code of HighCharts and adapt it to StationsRadar's specific needs.

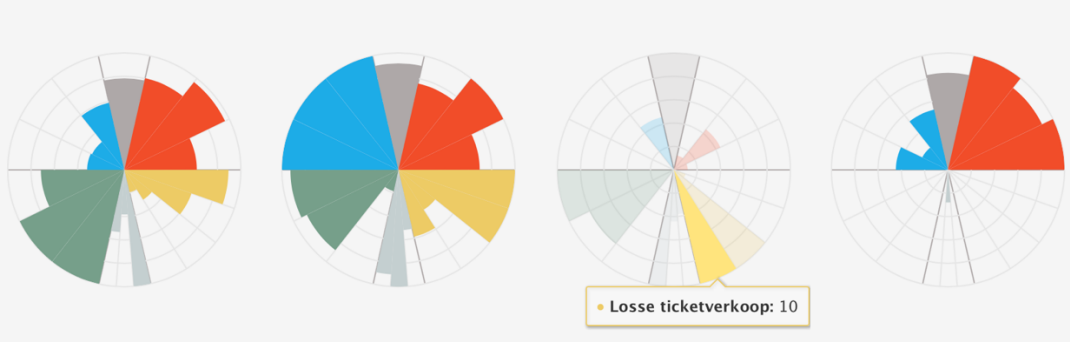


Figure 30 - screenshot of StationsRadar radar diagrams showing four stations (from left to right: Aalst, Bruges, Gent Dampoort, Brussels Airport).

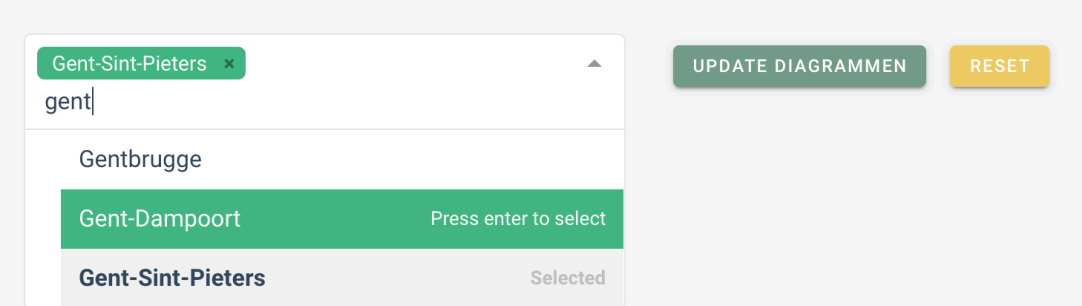


Figure 31 - StationsRadar radar diagram stations selection menu.

It is undoubtedly challenging to find the "right" balance between a data-driven tool accessible to non-technical people and a tool retaining enough similarities to a GIS application. This challenge forced me to look beyond the currently available solutions (e.g., HighCharts, R, Shiny), and incorporate all those solutions into an ecosystem capable of communicating data without losing its scientific value.

6.2 Limitations of current research

In this section, I will identify the limitations of the research presented in the formative chapters (Chapter 2 to Chapter 5) of this dissertation and discuss how these limitations have the potential to foster future avenues of research. When SKYNET was first developed, there were three core challenges present from the beginning: coding knowledge, continuous updates, and integration with other datasets. The first challenge pertained to my knowledge of R. When starting to develop SKYNET, my focus was on having the package able to parse the BTS DB1B and T-100 CSV files, and to have them in a format that would allow immediate analysis. As time progressed and as I developed my coding skills in R, I was able to implement some considerable changes in SKYNET, for example, adding the possibility to download the BTS data directly from SKYNET and some considerable improvements in terms of memory management and consistency. While most improvements were "under the hood"[95] changes, they allowed better data management (e.g., by allowing larger amounts of data to be analyzed), and better integration with other packages (e.g., by having a more consistent data output). These changes also allowed better continuity in terms of updates, as in programming, a more consistent language (i.e., coding) leads to better testing, debugging, and features updates.

In terms of continuous updates, the first challenge I faced was the CRAN requirements to constantly update the packages it hosts. While this is perfectly feasible in terms of workforce in the short run, in the longer run it becomes more challenging, as considerable efforts have to be put into software development, mostly because, in this case, there is only one developer. Regarding the data it uses, there are challenges as well in adding new air transport-related datasets. Currently SKYNET is only equipped to handle BTS DB1B and T-100 data. As mentioned at the end of Chapter 2, it would be of added value to incorporate different data sources (e.g., OAG, Flightaware, RDC Aviation). However, that would require access to those data sources, which in turn would be costly, and in case of changes to their APIs or to data structure, new access would have to be requested. The efforts and costs both in terms of time and money required to maintain, update, and upgrade SKYNET go beyond the resources currently available. A possible future avenue for research, therefore, would be to engage in a systematic collaboration with some of these data providers. Another option could be a collaboration with a research group that has structural access to these data sources, although this may not be ideal in the longer term as it does not ensure continuous access to data or API updates. Three years after it was first launched, SKYNET remains relevant as it is to the best of my knowledge the only open-source software capable of downloading air transport data whilst providing a set of analytical tools that allow its immediate analysis. However, time and resource constraints could be one of SKYNET's most critical limitations.

In Chapter 3, I presented a methodology to identify potential bias in the BTS DB1B dataset. The first challenge and limitation of this work pertains to the absence of detailed metadata describing collection and curation of both the DB1B and T-100 datasets. The only available metadata is buried in lengthy documents (i.e., 14 CFR 241,[96] Accounting and Reporting Directives[97]), which focus exclusively on guidelines and directives, without providing information on data quality and integrity. Unfortunately, our requests for more information led mostly to being directed to these two documents, without any further

[95] In programming, this term refers to changes not immediately visible to the user.
[96] https://www.govinfo.gov/app/details/CFR-2012-title14-vol4/CFR-2012-title14-vol4-part241
[97] https://www.bts.gov/topics/airlines-and-airports/accounting-and-reporting-directives

information being provided. This is an issue common to most online datasets. Either due to competition laws or arguments of data privacy, there are still considerable improvements needed to increase data transparency.

In Chapter 4 I identified several limitations to this research. First, I defined four distinct driving times (i.e., 20m, 30m, 45m, 60m) to each airport. These distinct driving times were defined to overcome computational and cost constraints (ArcGIS requires credits that have to be purchased in order to calculate OD matrices). However, the limitation in using time bands (i.e., instead of the exact driving time from each block group) leads to the aggregation of data, making it more difficult to understand fine-grained dynamics (e.g., the difference between block groups on the outer or inner border of time bands). Another challenge was that in order to simplify the description of MARs dynamics for the New York MAR, we set a limit of 60 minutes based on similar research. However, on a national scale, a limit of 60 minutes might not be necessarily realistic as has been demonstrated in previous research (Matisziw & Grubesic, 2010; T. H. Grubesic & Matisziw, 2011). Currently I am working on a project aimed to overcome most of these challenges. First, I was able to calculate the driving distance from each block group to every airport for a maximum 2 h, 30 m drive time. This should suffice to eliminate both the challenge of grouping block groups into time bands and the 60-minute limitation set in Chapter 4.

For this project I used three different air transport-related datasets from the BTS (i.e., DB1B, T-100, on-time performance). As mentioned before, one of the challenges in using these datasets was how they are temporally grouped: the DB1B is grouped per quarter, the T-100 per month. and the on-time performance per hour. There are some immediately obvious challenges in aggregating these datasets. For example, it is not possible to use air fares from the DB1B (grouped per quarter) to represent individual flights as seen in the on-time performance dataset (grouped per hour). This challenge led me to create the methodology described in detail in Chapter 4. That is, instead of aggregating all data, the algorithm is run individually for variables within the same temporal setting. For example, when comparing fares, we only use data from the DB1B dataset. However, for on-time performance we exclusively use the on-time performance dataset. After our algorithm runs the comparisons between routes and airports, an index is produced that is agnostic to time, and that allows us to use sources grouped in different temporal settings.

During the review process of Chapter 4, some reviewers pointed to the potentially limited relevance of the proposed methodology as it lacks proper calibration (e.g., against survey data). While it would be interesting to attune the output of our methodology against survey or other relevant data (e.g., social media), the goal was not to model stated preferences. Instead, I aimed to provide an index showing the best airport option given a set of preferences (i.e., fare, connectivity, on-time performance). Future research could focus on using different data sources (e.g., Twitter, UBER) to finetune this model or to incorporate AI solutions to better understand the dynamics previously observed.

I briefly discussed some of the challenges that emerged in Chapter 5, alongside possible avenues for future research. In this dissertation, I have mostly focused on the technical challenges and characteristics of StationsRadar. With that in mind, the most critical limitation when developing and maintaining StationsRadar concerns the development team consisting of only one person. This creates technical challenges both in time and skills needed to maintain, upgrade, and update this web-tool. First, the server that the tool is running from needs constant maintenance (e.g., software updates).

Second, in order to keep up with advances in technology, the tool's code should be updated, along with all the libraries it uses. Finally, as the data is exclusively managed by me and my colleague Freke Caset, we should in the longer term ensure that the data is up to date. In addition, it would also be interesting to bring functionalities to StationsRadar that would allow for a better interaction between users and data. For example, if there were a user portal, users could add or even flag issues with the data. However, this would require a development team to support the different requirements it would harbor (e.g., user authentication, backend development).

6.3 Final remarks

It is evident that we currently live in a world where data has become ubiquitous. The potential of harnessing this flood of data are no longer subject of fiction, with virtual personal assistants, self-driving cars and AI powered smartphones being or becoming a constant in our daily lives. We also have seen being used to change the outcome of elections (Isaak & Hanna, 2018; Ward, 2018) or to predict crime (Palantir, 2017; Winston, 2018). However, nowadays there is a high cost to processing the vast amounts of data being produced. For example, in the case of smart cities, sensors or smartphones can produce several petabytes[98] of data (M. Townsend, 2014), which requires costly and often complex computing power. Another challenge is that as data becomes more important, its economic value grows as well. For some, data has been often compared to oil (Hirsch, 2015) as it bears nearly unlimited potential. This in turn means that as data grows in value, we see the efforts in having a data democratization being hampered. Alongside with this flow of data and its impact in daily life and in research, the need for tools and visualisations aimed to simplify complex data rises (McCandless, 2014). While the field of geography does not harbour the answers to reducing data complexity, its multidisciplinarity holds some of the keys necessary to foster innovation. This means that as geographers, we have the necessary tools to assure that some of the challenges and restrictions mentioned throughout this dissertation will not curb the enthusiasm and potential of this new era of complex and ubiquitous data.

[98] 10^{15} bytes

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning. *Proc 12th USENIX Conference on Operating Systems Design and Implementation*.

Air Transport Action Group. (2010). The Economic and Social Benefits of Air Transport. In *Ovidius University Annals - Economic Sciences Series* (Issue 1).

Ameen, N., & Kamga, C. (2013). Forecast of airport ground access mode choice with the incremental logit model. *Transportation Research Record*, *2336*(2336), 97–104. https://doi.org/10.3141/2336-12

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*. https://doi.org/10.1016/j.ecolmodel.2009.09.008

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., Jern, M., Kraak, M. J., Schumann, H., & Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2010.508043

ArcGIS. (2019). *Historical traffic*. https://desktop.arcgis.com/en/arcmap/latest/extensions/network-analyst/traffic-historical-10-1-and-later.htm

Atelier Zuidvleugel. (2006). *Ruimte en Lijn. Ruimtelijke Verkenning Stedenbaan 2010 - 2020. Zuidvleugel van de Randstad* (Atelier Zuidvleugel (ed.)). http://www.atelierzuidvleugel.nl/files/Ruimte en Lijn_Ruimtelijke Verkenning Stedenbaan 2010-2020 I.pdf

Bagler, G. (2008). Analysis of the airport network of India as a complex weighted network. *Physica A: Statistical Mechanics and Its Applications*, *387*(12), 2972–2980. https://doi.org/10.1016/j.physa.2008.01.077

Balducci, A., & Bertolini, L. (2007). Reflecting on practice or reflecting with practice? *Planning Theory and Practice*. https://doi.org/10.1080/14649350701664770

Balz, V., & Schrijnen, J. (2009). From concept to projects: Stedenbaan, The Netherlands. In *Transit Oriented Development: Making it Happen* (pp. 75–90).

Bania, N., Bauer, P. W., & Zlatoper, T. J. (1998). U.S. air passenger service: A taxonomy of route networks, hub locations, and competition. *Transportation Research Part E: Logistics and Transportation Review*, *34*(1), 53–74. https://doi.org/10.1016/S1366-5545(97)00037-9

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2003). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3747–3752. https://doi.org/10.1073/pnas.0400087101

Batty, M. (2012). Smart Cities, Big Data. *Environment and Planning B: Planning and Design*, *39*(2), 191–193. https://doi.org/10.1068/b3902ed

Beck, A. T. (1993). Cognitive Therapy: Past, Present, and Future. *Journal of Consulting and Clinical Psychology*. https://doi.org/10.1037/0022-006X.61.2.194

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. In *Scientific American* (Vol. 284, Issue 5, pp. 34–43). https://doi.org/10.1038/scientificamerican0501-34

Bertolini, L. (1999). Spatial development patterns and public transport: The application of an

analytical model in the Netherlands. *Planning Practice and Research*. https://doi.org/10.1080/02697459915724

Bertolini, Luca. (2017). Planning the Mobile Metropolis. In *Planning the Mobile Metropolis* (pp. 1–14). Macmillan Education UK. https://doi.org/10.1057/978-1-137-31925-8_1

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.0610172104

Billger, M., Thuvander, L., & Wästberg, B. S. (2017). In search of visualization challenges: The development and implementation of visualization tools for supporting dialogue in urban planning processes. *Environment and Planning B: Urban Analytics and City Science*. https://doi.org/10.1177/0265813516657341

Blackstone, E. A., Buck, A. J., & Hakim, S. (2006). Determinants of airport choice in a multi-airport region. *Atlantic Economic Journal*, *34*(3), 313–326. https://doi.org/10.1007/s11293-006-9024-z

Bonnefoy, P. A., de Neufville, R., & Hansman, R. J. (2010). Evolution and development of multiairport systems: Worldwide perspective. *Journal of Transportation Engineering*, *136*(11), 1021–1029. https://doi.org/10.1061/(ASCE)0733-947X(2010)136:11(1021)

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. In *BioScience*. https://doi.org/10.1525/bio.2009.59.11.9

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. In *Science*. https://doi.org/10.1126/science.1251554

Borenstein, S., & Rose, N. L. (2002). Competition and Price Dispersion in the U.S. Airline Industry. *Journal of Political Economy*. https://doi.org/10.1086/261950

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895. https://doi.org/10.1126/science.1165821

Bounova, G. (2009). *Topological evolution of networks : case studies in the US airlines and language Wikipedias*. *2003*. http://dspace.mit.edu/handle/1721.1/62965

Boussauw, K., van Meeteren, M., Sansen, J., Meijers, E., Storme, T., Louw, E., Derudder, B., & Witlox, F. (2018). Planning for agglomeration economies in a polycentric region: Envisioning an efficient metropolitan core area in Flanders. *European Journal of Spatial Development*. https://doi.org/10.30689/EJSD2018:69.1650-9544

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*. https://doi.org/10.1080/1369118X.2012.678878

Brownstein, J. S., Wolfe, C. J., & Mandl, K. D. (2006). Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.0030401

Brueckner, J. K., Lee, D., & Singer, E. (2014). City-Pairs Versus Airport-Pairs: A Market-Definition Methodology for the Airline Industry. *Review of Industrial Organization*, *44*(1), 1–25. https://doi.org/10.1007/s11151-012-9371-7

Budd, T., Ison, S., & Ryley, T. (2011). Airport surface access in the UK: A management perspective. *Research in Transportation Business and Management*. https://doi.org/10.1016/j.rtbm.2011.05.003

Buneman, P., Khanna, S., & Tan, W. C. (2001). Why and where: A characterization of data provenance? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/3-540-

44503-x_20

Bureau of Transport Statistics. (2017). *A Time Series Analysis of Domestic Air Seat and Passenger Miles*. https://www.bts.gov/archive/publications/transportation_indicators/october_2002/Special/A_Time_Series_Analysis_of_Domestic_Air_Seat_and_Passenger_Miles

Bureau of Transport Statistics. (2018). *BTS DOT Airline Restricted Data*. https://www.bts.dot.gov/topics/airlines-and-airports/restricted-data

Bureau of Transport Statistics. (2019a). *Bureau of Transport Statistics*. https://www.bts.gov

Bureau of Transport Statistics. (2019b). *Certificated Air Carriers List*. https://www.transportation.gov/policy/aviation-policy/certificated-air-carriers-list

Bureau of Transport Statistics. (2019c). *Difference between T-100 Market and T-100 Segment Airline Traffic Data*. http://transportation.libanswers.com/faq/166158

Bureau of Transport Statistics. (2020). *Air Carrier Statistics (Form 41 Traffic)- All Carriers*. https://www.transtats.bts.gov/Tables.asp?DB_ID=111

Burghouwt, G., & de Wit, J. (2005). Temporal configurations of European airline networks. *Journal of Air Transport Management*, *11*(3), 185–198. https://doi.org/10.1016/j.jairtraman.2004.08.003

Button, K., & Lall, S. (1999). The economics of being an airport hub city. *Research in Transportation Economics*, *5*(C), 75–105. https://doi.org/10.1016/S0739-8859(99)80005-5

Button, K., & Yuan, J. (2013). Airfreight Transport and Economic Development: An Examination of Causality. *Urban Studies*, *50*(2), 329–340. https://doi.org/10.1177/0042098012446999

Campbell, J. F., & O'Kelly, M. E. (2012). Twenty-Five Years of Hub Location Research. *Transportation Science*, *46*(2), 153–169. https://doi.org/10.1287/trsc.1120.0410

Caset, F. (2019). *Planning for nodes, places, and people*. https://doi.org/http://hdl.handle.net/1854/LU-8637955

Caset, F., Teixeira, F. M., Derudder, B., Boussauw, K., & Witlox, F. (2019). Planning for nodes, places and people in Flanders and Brussels: Developing an empirical railway station assessment model for strategic decision-making. *Journal of Transport and Land Use*, *12*(1), 811–837. https://doi.org/10.5198/jtlu.2019.1483

Caset, F., Vale, D. S., & Viana, C. M. (2018). Correction to: Measuring the Accessibility of Railway Stations in the Brussels Regional Express Network: a Node-Place Modeling Approach. *Networks and Spatial Economics*. https://doi.org/10.1007/s11067-018-9409-y

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*. https://doi.org/10.1016/S1361-9209(97)00009-6

Champlin, C., te Brömmelstroet, M., & Pelzer, P. (2019). Tables, Tablets and Flexibility: Evaluating Planning Support System Performance under Different Conditions of Use. *Applied Spatial Analysis and Policy*, *12*(3), 467–491. https://doi.org/10.1007/s12061-018-9251-0

Chen, J. S. (2000). A Comparison of Information Usage Between Business and Leisure Travelers. *Journal of Hospitality & Leisure Marketing*, *7*(2), 65–76. https://doi.org/10.1300/J150v07n02_05

Cho, W., Windle, R. J., & Dresner, M. E. (2015). The impact of low-cost carriers on airport choice in the US: A case study of the Washington–Baltimore region. *Transportation Research Part E: Logistics and Transportation Review*, *81*, 141–157. https://doi.org/10.1016/j.tre.2015.06.004

Chollet François. (2015). Keras: The Python Deep Learning library. In *keras.io*. https://doi.org/10.1086/316861

Colizza, V., Barthélemy, M., Barrat, A., & Vespignani, A. (2007). Epidemic modeling in complex realities. In *Comptes Rendus - Biologies*. https://doi.org/10.1016/j.crvi.2007.02.014

Collobert, R., Van Der Maaten, L., & Joulin, A. (2016). Torchnet: An OpenSource Platform for (Deep) Learning Research. *Proceedings of the 33rd International Conference on Machine Learning*.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating "big data" and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*. https://doi.org/10.1080/15230406.2013.777137

CRAN. (2017). *CRAN - 10.000 packages*.

Dai, L., Derudder, B., & Liu, X. (2018). The evolving structure of the Southeast Asian air transport network through the lens of complex networks, 1979–2012. *Journal of Transport Geography*, *68*(April), 67–77. https://doi.org/10.1016/j.jtrangeo.2018.02.010

de Luca, S. (2012). Modelling airport choice behaviour for direct flights, connecting flights and different travel plans. *Journal of Transport Geography*, *22*, 148–163. https://doi.org/10.1016/j.jtrangeo.2011.12.006

de Neufville, R. (1995). Management of multi-airport systems. A development strategy. *Journal of Air Transport Management*, *2*(2), 99–110. https://doi.org/10.1016/0969-6997(95)00035-6

Debbage, K. G., & Delk, D. (2001). The geography of air passenger volume and local employment patterns by US metropolitan core area: 1973-1996. *Journal of Air Transport Management*, *7*(3), 159–167. https://doi.org/10.1016/S0969-6997(00)00045-4

Derudder, B., Devriendt, L., & Witlox, F. (2010). A spatial analysis of multiple airport cities. *Journal of Transport Geography*, *18*(3), 345–353. https://doi.org/10.1016/j.jtrangeo.2009.09.007

Derudder, B., & Witlox, F. (2005a). An appraisal of the use of airline data in assessing the world city network: A research note on data. *Urban Studies*, *42*(13), 2371–2388. https://doi.org/10.1080/00420980500379503

Derudder, B., & Witlox, F. (2005b). On the use of inadequate airline data in mappings of a global urban system. *Journal of Air Transport Management*, *11*(4), 231–237. https://doi.org/10.1016/j.jairtraman.2005.01.001

Dick, P. K. (1982). The Man in the High Castle. In *I Can*.

Distill. (2013). *Oakland International Airport branding and advertising*. http://wedistill.com/case-study/oakland-international-airport-branding-and-advertising/

Dobruszkes, F., & Van Hamme, G. (2011). The impact of the current economic crisis on the geography of air traffic volumes: An empirical analysis. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2011.07.015

Drakonakis, K., Ilia, P., Ioannidis, S., & Polakis, J. (2019). Please Forget Where I Was Last Summer: The Privacy Risks of Public Location (Meta)Data. *Proceedings 2019 Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2019.23151

Dresner, M., Lin, J. S. C., & Windle, R. (1996a). The impact of low-cost carriers on airport and route competition. *Journal of Transport Economics and Policy*, *30*(3), 309–328. https://doi.org/10.2307/20053709

Dresner, M., Lin, J. S. C., & Windle, R. (1996b). The impact of low-cost carriers on airport and route competition. *Journal of Transport Economics and Policy*, *30*(3), 309–328. https://doi.org/10.2307/20053709

Dresner, M., & Xu, K. (1995). Customer service, customer satisfaction, and corporate performance in the service sector. *Journal of Business Logistics*, *16*(1), 23. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Customer+service,+custom

er+satisfaction,+and+corporate+performance+in+the+service+sector#2

Ducruet, C., & Beauguitte, L. (2014). Spatial Science and Network Science: Review and Outcomes of a Complex Relationship. *Networks and Spatial Economics*, *14*(3–4), 297–316. https://doi.org/10.1007/s11067-013-9222-6

Ducruet, C., Rozenblat, C., & Zaidi, F. (2010). Ports in multi-level maritime networks: Evidence from the Atlantic (1996-2006). *Journal of Transport Geography*, *18*(4), 508–518. https://doi.org/10.1016/j.jtrangeo.2010.03.005

Duffhues, J., Mayer, I. S., Nefs, M., & Van Der Vliet, M. (2014). Breaking barriers to Transit-Oriented development: Insights from the serious game SPRINTCITY. *Environment and Planning B: Planning and Design*. https://doi.org/10.1068/b39130

Engin, Z., van Dijk, J., Lan, T., Longley, P. A., Treleaven, P., Batty, M., & Penn, A. (2020). Data-driven urban management: Mapping the landscape. *Journal of Urban Management*, *9*(2), 140–150. https://doi.org/10.1016/j.jum.2019.12.001

Entwisle, B., Rindfuss, R. R., Walsh, S. J., Evans, T. P., & Curran, S. R. (1997). Geographic information systems, spatial network analysis, and contraceptive choice. *Demography*. https://doi.org/10.2307/2061697

Ewing, R., & Cervero, R. (2010). Travel and the Built Environment: A Meta-Analysis. *Journal of the American Planning Association*. https://doi.org/10.1080/01944361003766766

Fahey, S. (2014). The Democratization of Big Data. *Journal of National Security Law & Policy*.

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. In *National Science Review*. https://doi.org/10.1093/nsr/nwt032

Floridi, L. (2012). Big data and their epistemological challenge. In *Philosophy and Technology*. https://doi.org/10.1007/s13347-012-0093-4

Franke, M. (2004). Competition between network carriers and low-cost carriers—retreat battle or breakthrough to a new level of efficiency? *Journal of Air Transport Management*, *10*(1), 15–21. https://doi.org/10.1016/j.jairtraman.2003.10.008

Freedman, C. (2013). Critical theory and science fiction. In *Choice Reviews Online* (Vol. 38, Issue 01). Wesleyan University Press. https://doi.org/10.5860/choice.38-0121

Frost, M. E., & Spence, N. A. (1995). The rediscovery of accessibility and economic potential: the critical issue of self-potential. *Environment & Planning A*, *27*(11), 1833–1848. https://doi.org/10.1068/a271833

Fu, Q., & Kim, A. M. (2016). Supply-and-demand models for exploring relationships between smaller airports and neighboring hub airports in the U.S. *Journal of Air Transport Management*, *52*(May), 67–79. https://doi.org/10.1016/j.jairtraman.2015.12.008

Fuellhart, K. (2007). Airport catchment and leakage in a multi-airport region: The case of Harrisburg International. *Journal of Transport Geography*, *15*(4), 231–244. https://doi.org/10.1016/j.jtrangeo.2006.08.001

Fuellhart, K., & O'Connor, K. (2019). A supply-side categorization of airports across global multiple-airport cities and regions. *GeoJournal*, *84*(1), 15–30. https://doi.org/10.1007/s10708-018-9847-6

Fuellhart, K., O'Connor, K., & Woltemade, C. (2013). Route-level passenger variation within three multi-airport regions in the USA. *Journal of Transport Geography*, *31*, 171–180. https://doi.org/10.1016/j.jtrangeo.2013.06.012

Fuellhart, K., Ooms, K., Derudder, B., & O'Connor, K. (2016). Patterns of US air transport across the economic unevenness of 2003–2013. *Journal of Maps*, *12*(5), 1253–1257. https://doi.org/10.1080/17445647.2016.1152917

Gallotti, R., Fuster, M., & Ramasco, J. J. (2017). New data sources to study airport competition. *SESAR Innovation Days*.

GAO. (1997). *Barriers to Entry Continue to Limit Benefits of Airline Deregulation*. https://doi.org/T-RCED-97-120

Goetz, A. R. (1993). Geographic patterns of air service frequencies and and pricing at US cities. *Journal of the Transportation Research Forum*, *33*, 56–72.

Goetz, A. R. (2002). Deregulation, competition, and antitrust implications in the US airline industry. *Journal of Transport Geography*, *10*(1), 1–19. https://doi.org/10.1016/S0966-6923(01)00034-5

Goetz, A. R., & Sutton, C. J. (1997). The geography of deregulation in the U.S. airline industry. *Annals of the Association of American Geographers*, *87*(2), 238–263. https://doi.org/10.1111/0004-5608.872052

Goetz, A. R., & Vowles, T. M. (2000). 'Pockets of Pain' across the deregulated landscape: the geography of US airline fares and service in the, 1990s. *Paper Presented at the Annual Meeting of the Association of American Geographers, Pittsburg*.

Goetz, A. R., & Vowles, T. M. (2009). A Hierarchical Typology of Intermodal Air-Rail Connections at Large Airports in the United States. *Association of American Geographers Annual Meeting*. http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=25865

Goodchild, M. F. (1991). Just the facts. *Political Geography Quarterly*. https://doi.org/10.1016/0260-9827(91)90001-B

Graham, M., & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography*. https://doi.org/10.1177/2043820613513121

Graham, M., & Zook, M. (2013). Augmented realities and uneven geographies: Exploring the geolinguistic contours of the web. *Environment and Planning A*. https://doi.org/10.1068/a44674

Graham, M., Zook, M., & Boulton, A. (2013). Augmented reality in urban places: Contested content and the duplicity of code. *Transactions of the Institute of British Geographers*. https://doi.org/10.1111/j.1475-5661.2012.00539.x

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., Dewitt, D. J., & Heber, G. (2005). Scientific data management in the coming decade. *SIGMOD Record*, *34*(4), 34–41. https://doi.org/10.1145/1107499.1107503

Groenendijk, L., Rezaei, J., & Correia, G. (2018). Incorporating the travellers' experience value in assessing the quality of transit nodes: A Rotterdam case study. *Case Studies on Transport Policy*. https://doi.org/10.1016/j.cstp.2018.07.007

Grubesic, T. H., & Matisziw, T. C. (2011). A spatial analysis of air transport access and the essential air service program in the United States. *Journal of Transport Geography*, *19*(1), 93–105. https://doi.org/10.1016/j.jtrangeo.2009.12.006

Grubesic, T., & Zook, M. (2007). A ticket to ride: Evolving landscapes of air travel accessibility in the United States. *Journal of Transport Geography*, *15*(6), 417–430. https://doi.org/10.1016/j.jtrangeo.2006.12.002

Guimera, R., Mossa, S., Turtschi, A., & Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, *102*(22), 7794–7799. https://doi.org/10.1073/pnas.0407994102

Gutiérrez, J. (2001). Location, economic potential and daily accessibility: An analysis of the accessibility impact of the high-speed line Madrid-Barcelona-French border. *Journal of Transport Geography*, *9*(4), 229–242. https://doi.org/10.1016/S0966-6923(01)00017-5

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*. https://doi.org/10.1068/b35097

Harris, R., Johnston, R., & Burgess, S. (2016). Tangled spaghetti: Modelling the core catchment areas of London's secondary schools. *Environment and Planning A: Economy and Space*, *48*(9), 1681–1683. https://doi.org/10.1177/0308518X15603987

Harrison, J., & Hoyler, M. (2014). Megaregions: Globalization's new urban form? In *Megaregions: Globalization's New Urban Form?* https://doi.org/10.4337/9781782547907

Harvey, G. (1987). Airport choice in a multiple airport region. *Transportation Research Part A: General*, *21*(6), 439–449. https://doi.org/10.1016/0191-2607(87)90033-1

Hayes, D. P. (1992). The growing inaccessibility of science. In *Nature*. https://doi.org/10.1038/356739a0

Heilman, J. (2017). *Spatial competition in airport markets : An application of the Huff model*.

Here. (2020). *Here Maps*. https://www.here.com

Hess, S. (2010). Evidence of passenger preferences for specific types of airports. *Journal of Air Transport Management*, *16*(4), 191–195. https://doi.org/10.1016/j.jairtraman.2009.11.006

Hess, S., Adler, T., & Polak, J. W. (2007). Modelling airport and airline choice behaviour with the use of stated preference survey data. *Transportation Research Part E: Logistics and Transportation Review*, *43*(3), 221–233. https://doi.org/10.1016/j.tre.2006.10.002

Hess, S., & Polak, J. W. (2005a). Mixed logit modelling of airport choice in multi-airport regions. *Journal of Air Transport Management*, *11*(2), 59–68. https://doi.org/10.1016/j.jairtraman.2004.09.001

Hess, S., & Polak, J. W. (2006). Airport, airline and access mode choice in the San Francisco Bay area. *Papers in Regional Science*, *85*(4), 543–567. https://doi.org/10.1111/j.1435-5957.2006.00097.x

Hess, S., & Polak, J. W. (2005b). Accounting for random taste heterogeneity in airport choice modeling. *Transportation Research Record*, *1915*, 36–43. https://doi.org/10.3141/1915-05

Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm. *Data-Intensive Scientific Discovery*. *Microsoft Research*.

Hirsch, D. D. (2015). The glass house effect: Big Data, the new oil, and the power of analogy. *Maine Law Review*, *66*(2), 374–395. http://heinonline.org

Hsiao, C.-Y., & Hansen, M. (2011). A passenger demand model for air transportation in a hub-and-spoke network. *Transportation Research Part E: Logistics and Transportation Review*, *47*(6), 1112–1125. https://doi.org/10.1016/j.tre.2011.05.012

Huang, Z., Wu, X., Garcia, A. J., Fik, T. J., & Tatem, A. J. (2013). An Open-Access Modeled Passenger Flow Matrix for the Global Air Network in 2010. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064317

Huff, D. L. (1963). A Probabilistic Analysis of Shopping Center Trade Areas. *Land Economics*, *39*(1), 81. https://doi.org/10.2307/3144521

Huff, D. L. (1964). Defining and Estimating a Trading Area. *Journal of Marketing*, *28*(3), 34. https://doi.org/10.2307/1249154

IATA. (2016). IATA Forecasts Passenger Demand to Double Over 20 Years. *IATA Press Release No.: 59*, *October*, 18–22. https://doi.org/10.1016/S0955-2219(03)00248-6

Ibraeva, A., Correia, G. H. de A., Silva, C., & Antunes, A. P. (2020). Transit-oriented development: A review of research achievements and challenges. *Transportation Research Part A: Policy and*

Practice. https://doi.org/10.1016/j.tra.2019.10.018

International Air Rail Organisation, & Blond, P. Le. (2013). *IARO report 17.13*. https://www.iaro.com/sitefiles/IARO Report 17.13.pdf

International Civil Aviation Organization. (2017). *Air transport, passengers carried*. https://data.worldbank.org/indicator/IS.AIR.PSGR?end=2017&start=1990

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*. https://doi.org/10.1109/MC.2018.3191268

Ishii, J., Jun, S., & Van Dender, K. (2009). Air travel choices in multi-airport markets. *Journal of Urban Economics*, *65*(2), 216–227. https://doi.org/10.1016/j.jue.2008.12.001

Ishutkina, M., & Hansman, R. J. (2008). Analysis of Interaction between Air Transportation and Economic Activity. *The 26th Congress of ICAS and 8th AIAA ATIO*. https://doi.org/10.2514/6.2008-8888

Jacquez, G. M. (2008). Spatial Cluster Analysis. In *The Handbook of Geographic Information Science*. https://doi.org/10.1002/9780470690819.ch22

Jonah, L. (2010). A Physicist Solves the City. *New York Times Magazine*.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual analytics: Scope and challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-540-71080-6_6

Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). Challenges in visual data analysis. *Proceedings of the International Conference on Information Visualisation*. https://doi.org/10.1109/IV.2006.31

Keim Daniel, K. J., & Mansmann, G. rey E. and F. (2010). Mastering the Information Age Solving Problems with Visual Analytics. In *Mastering the Information Age Solving Problems with Visual Analytics*. https://doi.org/10.1016/j.procs.2011.12.035

Kickert, C. C., Pont, M. B., & Nefs, M. (2014). Surveying density, urban characteristics, and development capacity of station areas in the delta metropolis. *Environment and Planning B: Planning and Design*. https://doi.org/10.1068/b39020

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*. https://doi.org/10.1177/2043820613513388

Kitchin, R. (2014). The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences. In *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. SAGE Publications Ltd. https://doi.org/10.4135/9781473909472

Kolb, D. A., & Fry, R. E. (1974). Toward an Applied Theory of Experiential Learning. In *Theories of Group Process*.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Koster, P., Kroes, E., & Verhoef, E. (2011). Travel time variability and airport accessibility. *Transportation Research Part B: Methodological*, *45*(10), 1545–1559. https://doi.org/10.1016/j.trb.2011.05.027

Lassen, C. (2006). Aeromobility and work. *Environment and Planning A*, *38*(2), 301–312. https://doi.org/10.1068/a37278

Levine, M. E. (1987). Airline Competition in Deregulated Markets : Theory , Firm Strategy , and Public Policy. *Yale Journal on Regulation*, *4*(2), 393–494.

Li, W., & Cai, X. (2004). Statistical analysis of airport network of China. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *69*(4), 6. https://doi.org/10.1103/PhysRevE.69.046106

Lian, J. I., & Rønnevik, J. (2011). Airport competition - Regional airports losing ground to main airports. *Journal of Transport Geography*, *19*(1), 85–92. https://doi.org/10.1016/j.jtrangeo.2009.12.004

Lieshout, R. (2012). Measuring the size of an airport's catchment area. *Journal of Transport Geography*, *25*, 27–34. https://doi.org/10.1016/j.jtrangeo.2012.07.004

Lieshout, R., Malighetti, P., Redondi, R., & Burghouwt, G. (2016). The competitive landscape of air transport in Europe. *Journal of Transport Geography*, *50*, 68–82. https://doi.org/10.1016/j.jtrangeo.2015.06.001

Lin, W. (2014). The politics of flying: Aeromobile frictions in a mobile city. *Journal of Transport Geography*, *38*, 92–99. https://doi.org/10.1016/j.jtrangeo.2014.06.002

Liu, X., Derudder, B., & Wu, K. (2016). Measuring Polycentric Urban Development in China: An Intercity Transportation Network Perspective. *Regional Studies*, *50*(8), 1302–1315. https://doi.org/10.1080/00343404.2015.1004535

Lium, A.-G., Crainic, T. G., & Wallace, S. W. (2009). A Study of Demand Stochasticity in Service Network Design. *Transportation Science*, *43*(2), 144–157. https://doi.org/10.1287/trsc.1090.0265

Lohr, S. (2012). Sure , Big Data Is Great . But So Is Intuition. *The New York Times*.

Lohr, S. (2013). Sizing Up Big Data, Broadening Beyond the Internet. *The New York Times*.

Lohr, S. (2014). The Age of Big Data. *Science*. https://doi.org/10.1126/science.1243089

Loo, B. P. Y. (2008). Passengers' airport choice within multi-airport regions (MARs): some insights from a stated preference survey at Hong Kong International Airport. *Journal of Transport Geography*, *16*(2), 117–125. https://doi.org/10.1016/j.jtrangeo.2007.05.003

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Luo, D. (2014). The Price Effects of the Delta/Northwest Airline Merger. *Review of Industrial Organization*. https://doi.org/10.1007/s11151-013-9380-1

Luttmann, A. (2019). Capacity Constraints and Service Quality: Do Airport Slot Controls Reduce Flight Delays? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3386866

M. Townsend, A. (2014). Smart cities: big data, civic hackers, and the quest for a new utopia. *Choice Reviews Online*, *51*(06), 51-3557-51–3557. https://doi.org/10.5860/CHOICE.51-3557

Mainzer, B. W. (2007). Global Airlines: Competition in a Transnational Industry. *Journal of Revenue and Pricing Management*, *6*(2), 155–156. https://doi.org/10.1057/palgrave.rpm.5160078

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population bias in geotagged tweets. *AAAI Workshop - Technical Report*.

Mandel, B. (2014). Contemporary Airport Demand Forecasting Choice models and air transport forecasting. *OECD International Transport Forum, Discussion Paper No. 2014-07*, *07*, 35. http://www.internationaltransportforum.org/jtrc/DiscussionPapers/DP201407.pdf

Mao, L., Wu, X., Huang, Z., & Tatem, A. J. (2015). Modeling monthly flows of global air travel passengers: An open-access data resource. *Journal of Transport Geography*, *48*(October), 52–60. https://doi.org/10.1016/j.jtrangeo.2015.08.017

Marcucci, E., & Gatta, V. (2011). Regional airport choice: Consumer behaviour and policy implications. *Journal of Transport Geography*, *19*(1), 70–84. https://doi.org/10.1016/j.jtrangeo.2009.10.001

Marques Teixeira, F., & Derudder, B. (2020). Revealing route bias in air transport data: The case of the Bureau of Transport Statistics (BTS), Origin-Destination Survey (DB1B). *Journal of Air Transport Management*, *82*(October 2019), 101745. https://doi.org/10.1016/j.jairtraman.2019.101745

Matisziw, T. C., & Grubesic, T. H. (2010). Evaluating locational accessibility to the US air transportation system. *Transportation Research Part A: Policy and Practice*, *44*(9), 710–722. https://doi.org/10.1016/j.tra.2010.07.004

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McCandless, D. (2014). *Knowledge is Beautiful*. https://books.google.co.uk/books/about/Knowledge_is_Beautiful.html?id=ZpQ6YgEACAAJ&pgis=1

Merton, R. K. (2007). On Sociological Theories of the Middle Range [1949]. *Classical Sociological Theory*. https://doi.org/10.1007/3-540-27354-9_1

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*. https://doi.org/10.1007/s10708-014-9602-6

Monger, J. I. (2010). Thirsting for equal protection: The legal implications of municipal water access in kennedy v. city of zanesville and the need for federal oversight of governments practicing unlawful race discrimination. *Catholic University Law Review*.

Morrison, S. A., & Winston, C. (1990). The dynamics of airline pricing and competition. *American Economic Review*.

Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased? Assessing the representativeness of twitter's streaming API. *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web*, 555–556. https://doi.org/10.1145/2567948.2576952

Mun, S. Il, & Teraji, Y. (2012). The organisation of multiple airports in a metropolitan area. *Journal of Transport Economics and Policy*, *46*(2), 221–237.

Muñoz, C., Cordoba, J., & Sarmiento, I. (2017). Airport choice model in multiple airport regions. *Journal of Airline and Airport Management*, *7*(1), 1. https://doi.org/10.3926/jairm.62

Neal, Z. (2010). Refining the air traffic approach to city networks. *Urban Studies*, *47*(10), 2195–2215. https://doi.org/10.1177/0042098009357352

Neal, Z. (2014a). AIRNET: A Programme for Generating Intercity Networks. *Urban Studies*, *51*(1), 136–152. https://doi.org/10.1177/0042098013484537

Neal, Z. (2014b). The devil is in the details: Differences in air traffic networks by scale, species, and season. *Social Networks*, *38*(1), 63–73. https://doi.org/10.1016/j.socnet.2014.03.003

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Nigro, A., Bertolini, L., & Moccia, F. D. (2019). Land use and public transport integration in small cities and towns: Assessment methodology and application. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2018.11.004

Notaerts, L., Meganck, R., Inslegers, R., & Krivzov, J. (2017). Beyond Clinical Case Studies in Psychoanalysis: A Review of Psychoanalytic Empirical Single Case Studies Published in ISI-

Ranked Journals. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01749

O'Connor, K., & Fuellhart, K. (2012). Cities and air services: The influence of the airline industry. *Journal of Transport Geography*, *22*(April 2011), 46–52. https://doi.org/10.1016/j.jtrangeo.2011.10.007

O'Connor, K., & Fuellhart, K. (2016). Airports and regional air transport markets: A new perspective. *Journal of Transport Geography*, *53*, 78–82. https://doi.org/10.1016/j.jtrangeo.2015.10.010

O'Kelly, M. E. (1998). A geographer's analysis of hub-and-spoke networks. *Journal of Transport Geography*, *6*(3), 171–186. https://doi.org/10.1016/S0966-6923(98)00010-6

O'Kelly, M. E. (2016). Global Airline Networks: Comparative Nodal Access Measures. *Spatial Economic Analysis*, *11*(3), 253–275. https://doi.org/10.1080/17421772.2016.1177262

O'Kelly, M. E., Bryan, D., Skorin-Kapov, D., & Skorin-Kapov, J. (1996). Hub network design with single and multiple allocation: A computational study. *Location Science*, *4*(3), 125–138. https://doi.org/10.1016/S0966-8349(96)00015-0

O'Kelly, M. E., & Miller, H. J. (1994). The hub network design problem. A review and synthesis. *Journal of Transport Geography*, *2*(1), 31–40. https://doi.org/10.1016/0966-6923(94)90032-9

O'Sullivan, D., & Manson, S. M. (2015). Do Physicists Have Geography Envy? And What Can Geographers Learn from It? *Annals of the Association of American Geographers*, *105*(4), 704–722. https://doi.org/10.1080/00045608.2015.1039105

OAG. (2015). *DOT Analyser User Guide*. https://www.oag.com/dot-analyser-user-guide#1.3.2

OAG. (2019). *Official Aviation Guide*. www.oag.com

Offenhuber, D., Ratti, C., Szell, M., & Groß, B. (2014). Hubcab – Exploring the Benefits of Shared Taxi Services. In *Decoding the City*. https://doi.org/10.1515/9783038213925.28

Office of the Secretary - Department of Transportation. (2019). *14 CFR Part 241*. https://www.govinfo.gov/content/pkg/CFR-2012-title14-vol4/pdf/CFR-2012-title14-vol4-part241.pdf

Openshaw, S. (1991). A View on the GIS Crisis in Geography, or, Using GIS to Put Humpty-Dumpty Back Together Again. *Environment and Planning A: Economy and Space*. https://doi.org/10.1068/a230621

Openshaw, S. (1997). The truth about ground truth. *Transactions in GIS*, *2*(1), 7–24. https://doi.org/10.1111/j.1467-9671.1997.tb00002.x

Ostrowski, P. L., O'Brien, T. V., & Gordon, G. L. (1993). Service Quality and Customer Loyalty in the Commercial Airline Industry. *Journal of Travel Research*, *32*(2), 16–24. https://doi.org/10.1177/004728759303200203

Palantir. (2017). *Crime risk forecasting* (Patent No. US9836694B2). https://patents.google.com/patent/US9836694B2/en

Paliska, D., Drobne, S., Borruso, G., Gardina, M., & Fabjan, D. (2016). Passengers' airport choice and airports' catchment area analysis in cross-border Upper Adriatic multi-airport region. *Journal of Air Transport Management*, *57*, 143–154. https://doi.org/10.1016/j.jairtraman.2016.07.011

Papa, E., Carpentieri, G., & Angiello, G. (2018). A TOD classification of metro stations: An application in Naples. In *Green Energy and Technology*. https://doi.org/10.1007/978-3-319-77682-8_17

Papa, E., te Brömmelstroet, M., Silva, C., & Hull, A. (2016). Accessibility instruments for planning practice: A review of European experiences. In *Journal of Transport and Land Use*. https://doi.org/10.5198/jtlu.2015.585

Park, Y., & O'Kelly, M. E. (2016). Origin–destination synthesis for aviation network data: examining

hub operations in the domestic and international US markets. *Journal of Advanced Transportation*, *50*(8), 2288–2305. https://doi.org/10.1002/atr.1459

Patil, P., Peng, R., & Leek, J. (2016). A statistical definition for reproducibility and replicability. *BioRxiV*. https://doi.org/10.1101/066803

Pels, E., Nijkamp, P., & Rietveld, P. (2000). Airport and Airline Competition for Passengers Departing from a Large Metropolitan Area. *Journal of Urban Economics*, *48*(1), 29–45. https://doi.org/10.1006/juec.1999.2156

Pels, E., Nijkamp, P., & Rietveld, P. (2001). Airport and airline choice in a multiple airport region: An empirical analysis for the San Francisco bay area. *Regional Studies*, *35*(1), 1–9. https://doi.org/10.1080/00343400120025637

Pels, E., Nijkamp, P., & Rietveld, P. (2003). Access to and competition between airports: A case study for the San Francisco Bay area. *Transportation Research Part A: Policy and Practice*, *37*(1), 71–83. https://doi.org/10.1016/S0965-8564(02)00007-1

Pelzer, P. (2017). Usefulness of planning support systems: A conceptual framework and an empirical illustration. *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/j.tra.2016.06.019

Pitfield, D. E. (2008). The Southwest effect: A time-series analysis on passengers carried by selected routes and a market share comparison. *Journal of Air Transport Management*, *14*(3), 113–122. https://doi.org/10.1016/j.jairtraman.2008.02.006

Polidoro, F., Giannini, R., Conte, R. Lo, Mosca, S., & Rossetti, F. (2015). Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. In *Statistical Journal of the IAOS*. https://doi.org/10.3233/sji-150901

Poorthuis, A., & Zook, M. (2014). Artists and Bankers and Hipsters, Oh My! Mapping Tweets in the New York Metropolitan Region. *Cityscape*, *16*(2), 169–172. http://www.jstor.org/stable/26326893

Poorthuis, A., & Zook, M. (2015). Small Stories in Big Data : Gaining Insights From Large Spatial Point Pattern Datasets. *Cityscape: A Journal of Policy Development and Research*, *17*(1), 151–160. https://www.jstor.org/stable/26326929

Poorthuis, A., & Zook, M. (2017). Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. *Journal of Urban Technology*, *24*(4), 115–135. https://doi.org/10.1080/10630732.2017.1335153

Poorthuis, A., Zook, M., Shelton, T., Graham, M., & Stephens, M. (2014). Using Geotagged Digital Social Data in Geographic Research. *Key Methods in Geography*.

Porter, T. M. (2009). How Science Became Technical. *Isis*, *100*(2), 292–309. https://doi.org/10.1086/599552

Province of North Holland, & Deltametropolis Association. (2013). *Maak Plaats! Werken aan knooppuntontwikkeling in Noord-Holland*. (C. Edens (ed.)). Provincie Noord Holland.

R Team, C. (2017). R Core Team (2017). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria. URL Http://Www. R-Project. Org/., Page R Foundation for Statistical Computing*.

Rocha, L. E. C. (2017). Dynamics of air transport networks: A review from a complex systems perspective. *Chinese Journal of Aeronautics*, *30*(2), 469–478. https://doi.org/10.1016/j.cja.2016.12.029

Roucolle, C., Seregina, T., & Urdanoz, M. (2020). Measuring the development of airline networks: Comprehensive indicators. *Transportation Research Part A: Policy and Practice*, *133*, 303–324. https://doi.org/10.1016/j.tra.2019.12.010

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. https://doi.org/10.1126/science.346.6213.1063

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S. H., & Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, *111*(37), 13290–13294. https://doi.org/10.1073/pnas.1403657111

Scheurer, J., Curtis, C., & Porta, S. (2009). Spatial Network Analysis of Public Transport Systems: Developing a Strategic Planning Tool to Assess the Congruence of Movement and Urban Structure in Australian Cities. *4th State of Australian Cities Conference*, *July 2015*, 1–23.

Schuurman, N. (2000). Trouble in the heartland: GIS and its critics in the 1990s. *Progress in Human Geography*. https://doi.org/10.1191/030913200100189111

Seshadri, A., Baik, H., & Trani, A. (2007). *A Model to Estimate Origin-Transfer-Destination Route Flows and Origin-Destination Segment Flows across the Continental United States*. *540*.

Silva, Cecília, Bertolini, L., te Brömmelstroet, M., Milakis, D., & Papa, E. (2017). Accessibility instruments in planning practice: Bridging the implementation gap. *Transport Policy*, *53*, 135–145. https://doi.org/10.1016/j.tranpol.2016.09.006

Silva, Cecília, & Larsson, A. (2018). Challenges for accessibility planning and research in the context of sustainable mobility - Discussion Paper. *International Transport Forum*.

Silva, Cecilia, Pinto, N., & Bertolini, L. (2019). *Designing Accessibility instruments: Lessons on Their Usability for integrated Land Use and Transport Planning Practices*.

Silvertown, J. (2009). A new dawn for citizen science. In *Trends in Ecology and Evolution*. https://doi.org/10.1016/j.tree.2009.03.017

Singh, Y. J., Lukman, A., Flacke, J., Zuidgeest, M., & Van Maarseveen, M. F. A. M. (2017). Measuring TOD around transit nodes - Towards TOD policy. *Transport Policy*, *56*, 96–111. https://doi.org/10.1016/j.tranpol.2017.03.013

Skinner, R. E. (1976). Airport choice: An empirical study. *Transportation Engineering Journal of ASCE*, *102*(4), 871–882.

Smith, D. A., & Timberlake, M. F. (2001). World city networks and hierarchies, 1977-1997: An empirical analysis of global air travel links. *American Behavioral Scientist*, *44*(10), 1656–1678. https://doi.org/10.1177/00027640121958104

Smith, N. (1992). History and philosophy of geography: real wars, theory wars. *Progress in Human Geography*, *16*(2), 257–271. https://doi.org/10.1177/030913259201600208

Soler, J. M., Cuartero, F., & Roblizo, M. (2012). Twitter as a tool for predicting elections results. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*. https://doi.org/10.1109/ASONAM.2012.206

Song, M. G., & Yeo, G. T. (2017). Analysis of the Air Transport Network Characteristics of Major Airports. *Asian Journal of Shipping and Logistics*, *33*(3), 117–125. https://doi.org/10.1016/j.ajsl.2017.09.002

Stadsregio Arnhem Nijmegen. (2011). *Knooppunten! Bereikbaarheid en ruimtelijke ontwikkeling op knooppunten van openbaar vervoer*.

Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M., Zdonik, S., Pagan, A., & Xu, S. (2013). Data curation at scale: The data tamer system. *CIDR 2013 - 6th Biennial Conference on Innovative Data Systems Research*.

Straatemeier, T. (2019). *Joint Accessibility Design: A framework to improve integrated transport and land use strategy making* [University of Amsterdam]. https://hdl.handle.net/11245.1/627ed142-4550-45a1-a1d1-ab1216be324d

Straatemeier, T., Bertolini, L., te Brömmelstroet, M., & Hoetjes, P. (2010). An experiential approach to

research in planning. *Environment and Planning B: Planning and Design*. https://doi.org/10.1068/b35122

Strohmeier, M., Martinovic, I., Fuchs, M., Schäfer, M., & Lenders, V. (2015). OpenSky: A swiss army knife for air traffic security research. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, 4A11-4A114. https://doi.org/10.1109/DASC.2015.7311411

Sun Country Airlines. (2018). *Sun Country Airlines*. https://www.suncountry.com/booking/search.html

Sun, X., Wandelt, S., & Linke, F. (2015). Temporal evolution analysis of the European air transportation system: Air navigation route network and airport network. *Transportmetrica B*, *3*(2), 153–168. https://doi.org/10.1080/21680566.2014.960504

Sun, X., Wandelt, S., & Zhang, A. (2017). A High-Resolution, Yet Scalable Framework for Transport Infrastructure Accessibility Based on Open Big Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3033166

Suzuki, Y. (2000). The relationship between on-time performance and airline market share: A new approach. *Transportation Research Part E: Logistics and Transportation Review*, *36*(2), 139–154. https://doi.org/10.1016/S1366-5545(99)00026-5

Suzuki, Y., & Audino, M. J. (2003). The effect of airfares on airport leakage in single-airport regions. *Transportation Journal*, *42*(5), 31–41.

Suzuki, Y., & Tyworth, J. E. (1998). A theoretical framework for modeling sales-service relationships in the transportation industry. *Transportation Research Part E: Logistics and Transportation Review*, *34*(2), 87–100. https://doi.org/10.1016/S1366-5545(98)00005-2

Szell, M., Grauwin, S., & Ratti, C. (2014). Contraction of online response to major events. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0089052

Taaffe, E. J. (1956). Air Transportation and United States Urban Distribution. *Geographical Review*, *46*(2), 219–238. https://doi.org/10.2307/211645

Taylor, P. J. (1990). Editorial comment GKS. In *Political Geography Quarterly*. https://doi.org/10.1016/0260-9827(90)90023-4

te Brommelstroet, M. C. G., Silva, C., Bertolini, L., & others. (2014). COST Action TU1002-Assessing usability of accessibility instruments. In *CITTA 7th Annual Conference*.

Teixeira, F., & Derudder, B. (2018). SKYNET: An R package for generating air passenger networks for urban studies. *Urban Studies*, *56*(14), 3030–3044. https://doi.org/10.1177/0042098018803258

Thatcher, J., Bergmann, L., Ricker, B., Rose-Redwood, R., O'Sullivan, D., Barnes, T. J., Barnesmoore, L. R., Beltz Imaoka, L., Burns, R., Cinnamon, J., Dalton, C. M., Davis, C., Dunn, S., Harvey, F., Jung, J. K., Kersten, E., Knigge, L. D., Lally, N., Lin, W., … Young, J. C. (2015). Revisiting critical GIS. *Environment and Planning A*. https://doi.org/10.1177/0308518X15622208

Thelle, M. H., & Sonne, M. la C. (2018). Airport competition in Europe. *Journal of Air Transport Management*, *67*, 232–240. https://doi.org/10.1016/j.jairtraman.2017.03.005

Uriely, N. (2001). "Travelling workers" and "working tourists": variations across the interaction between work and tourism. *International Journal of Tourism Research*, *3*(1), 1–8. https://doi.org/10.1002/1522-1970(200101/02)3:1<1::AID-JTR241>3.0.CO;2-M

US Census Bureau. (2017). *New York Metropolitan Area*.

US Census Bureau. (2018). *Census Blocks and Block Groups*. https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf

US Department of Transportation. (2012). *14 CFR 241 section 19-7 - Passenger Origin-Destination Survey*. https://www.govinfo.gov/app/details/CFR-2012-title14-vol4/CFR-2012-title14-vol4-

part241

US Department of Transportation. (2019). *DOT "on-time" performance*. https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120&DB_URL=

US Department of Transportation, & Office of the Secretary. (2005). *Aviation Data Modernization*. https://www.ecfr.gov/cgi-bin/text-idx?SID=8c512bbe8b110228dfb4f6dce7b5139e&mc=true&node=pt14.4.241&rgn=div5#se14.4.241_119_67

Vale, D. S., Viana, C. M., & Pereira, M. (2018). The extended node-place model at the local scale: Evaluating the integration of land use and transport for Lisbon's subway network. *Journal of Transport Geography*. https://doi.org/10.1016/j.jtrangeo.2018.05.004

Varga, M., & Varga, C. (2016). *Visual Analytics: Data, Analytical and Reasoning Provenance* (pp. 141–150). https://doi.org/10.1007/978-3-319-40226-0_9

Vowles, T. M. (2001). The "Southwest Effect" in multi-airport regions. *Journal of Air Transport Management*, *7*(4), 251–258. https://doi.org/10.1016/S0969-6997(01)00013-8

Vowles, T. M. (2006). Geographic perspectives of air transportation. *Professional Geographer*, *58*(1), 12–19. https://doi.org/10.1111/j.1467-9272.2006.00508.x

Ward, K. (2018). Social networks, the 2016 US presidential election, and Kantian ethics: applying the categorical imperative to Cambridge Analytica's behavioral microtargeting. *Journal of Media Ethics*, *33*(3), 133–148. https://doi.org/10.1080/23736992.2018.1477047

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Wei, F., & Grubesic, T. (2016). The pain persists: Exploring the spatiotemporal trends in air fares and itinerary pricing in the United States, 2002–2013. *Journal of Air Transport Management*, *57*, 107–121. https://doi.org/10.1016/j.jairtraman.2016.07.018

Wickham, C. (2011). A tale of two Airports: Exploring flight Traffic at SFO and OAK. *Journal of Computational and Graphical Statistics*, *20*(2), 291–293. https://doi.org/10.1198/jcgs.2011.4de

WINDLE, R., & DRESNER, M. (1995). The short and long run effects of entry on U.S. domestic air routes. *Transportation Journal*, *35*(2), 14–25.

Winston, A. (2018). *Palantir has secretly been using New Orleans to test its predictive policing technology*. The Verge. https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd

Wu, C., Han, J., & Hayashi, Y. (2011). Airport attractiveness analysis through a gravity model: a case study of Chubu international airport in Japan. *Proceedings of the Eastern Asia Society for Transportation Studies 2011,CD-ROM, Juju, Korea*, *8*(8).

Xu, Z., & Harriss, R. (2008). Exploring the structure of the U.S. intercity passenger air transportation network: A weighted complex network approach. *GeoJournal*, *73*(2), 87–102. https://doi.org/10.1007/s10708-008-9173-5

Yang, Z., Yu, S., & Notteboom, T. (2016). Airport location in multiple airport regions (MARs): The role of land and airside accessibility. *Journal of Transport Geography*, *52*, 98–110. https://doi.org/10.1016/j.jtrangeo.2016.03.007

Yazici, M. A., Kamga, C., & Singhal, A. (2013). A big data driven model for taxi drivers' airport pick-up decisions in New York City. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, 37–44. https://doi.org/10.1109/BigData.2013.6691775

Yeh, A. G.-O., & Chen, Z. (2020). From cities to super mega city regions in China in a new wave of urbanisation and economic transition: Issues and challenges. *Urban Studies*, *57*(3), 636–654.

https://doi.org/10.1177/0042098019879566

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2012). Quality Assessment Methodologies for Linked Open Data. *Semantic Web – Interoperability, Usability, Applicability*.

Zhang, Y., & Xie, Y. (2005). Small community airport choice behavior analysis: A case study of GTR. *Journal of Air Transport Management*, *11*(6), 442–447. https://doi.org/10.1016/j.jairtraman.2005.07.008

Zhong, C., Arisona, S. M., Huang, X., Batty, M., & Schmitt, G. (2014). Detecting the dynamics of urban structure through spatial network analysis. *International Journal of Geographical Information Science*, *28*(11), 2178–2199. https://doi.org/10.1080/13658816.2014.914521

Zhou, H., Xia, J. (Cecilia), Luo, Q., Nikolova, G., Sun, J., Hughes, B., Kelobonye, K., Wang, H., & Falkmer, T. (2018). Investigating the impact of catchment areas of airports on estimating air travel demand: A case study of regional Western Australia. *Journal of Air Transport Management*, *70*, 91–103. https://doi.org/10.1016/j.jairtraman.2018.05.001

Summary

This dissertation deals with the need for, and the key role and development of data-driven tools and methodologies for minimizing data complexity in transport geography research. It does so by developing a diverse range of analytical tools that collectively show that this is both feasible and useful. In addition to developing these tools in the strict sense, the dissertation also examines various challenges associated with the process, ranging from acquiring data to data curation, analysis and visualization. The thesis is divided into four chapters: (1) I first introduce SKYNET, a flexible R package that allows generating bespoke air transport statistics for urban studies based on publicly available data from the BTS in the United States. (2) I explore how potential biases in air transport datasets can be revealed and detailed by focusing on the US Origin- Destination Survey (DB1B) and the Air Carrier Statistics-form 41 traffic (T-100) datasets. (3) I explore the spatio-temporal dynamics of airport catchment areas within the New York Multi Airport Region. (4) And finally, I present StationsRadar, a data-driven web-based tool developed to support integrated land use and transport strategy-making at railway station locations in the region of Flanders and the Brussels Capital Region.

Samenvatting

Dit proefschrift focust op de nood aan, de sleutelrol voor, en de ontwikkeling van datagestuurde tools en methodologieën voor het aanpakken van datacomplexiteit in transportgeografisch onderzoek. Deze doelstelling wordt bereikt door het ontwikkelen van een aantal analytische tools die collectief aantonen dat dit zowel haalbaar als nuttig is. Naast het ontwikkelen van de tools in strikte zin, onderzoekt het proefschrift ook verschillende uitdagingen die met de procesontwikkeling ervan samenhangen, gaande van dataverwerving en -organisatie tot data-analyse en -visualisatie. Het proefschrift is onderverdeeld in vier hoofdstukken: (1) Ik introduceer SKYNET, een flexibel R-pakket waarmee op maat gemaakte luchtvaartstatistieken kunnen worden gegenereerd voor stedelijke studies, en dit op basis van openbare gegevens van het BTS in de Verenigde Staten. (2) Ik onderzoek hoe mogelijke verstoringseffecten in luchtvaartdatasets kunnen worden geïdentificeerd op basis van een analyse van de US Origin-Destination Survey (DB1B) en de Air Carrier Statistics-form 41 traffic (T-100)-datasets. (3) Ik onderzoek de tijd-ruimtelijke dynamiek in de ommelanden van luchthavens binnen de New York 'Multi Airport Region'. (4) Tot slot stel ik StationsRadar voor, een datagestuurde en webgebaseerde tool die werd ontwikkeld ter ondersteuning van geïntegreerde strategieën voor landgebruik en transport op stationslocaties in het Vlaams Gewest en het Brussels Hoofdstedelijk Gewest.

About the author

Filipe Alberto Marques Teixeira (°1984) is a Portuguese Biochemist (MSc) educated at the University of Coimbra, Portugal. After working in Germany as an Architect in 2009, he would start his studies in Biochemistry at the University of Coimbra where he would graduate in 2011, with a specialisation in Neurobiology. His main expertise and focus was on the effects of cannabinoids in the prefrontal cortex and their relationship with neuropsychiatric disorders. Between 2011 and 2012 he traveled around the world, to later move to Belgium in 2012. He left academia in 2012 to work for big pharma, as Service Support Manager until 2016. In 2017 Filipe started a PhD in Geography, more specifically in air transport geography. Later in the PhD the main focus steered to the development of data-driven tools to minimise data complexity in transport geography research. His research interests revolve around R programming, network analysis, machine learning, and visualisation of complex data.
Currently he combines his PhD with a position as a FLAMES consultant for R, spatial analysis and big data.


@filipeabroad


https://github.com/FilipeamTeixeira


https://www.researchgate.net/profile/Filipe_Alberto_Marques_Teixeira

Scholarly publications

Marques Teixeira, F. & Derudder, B. (2021)
Spatio-temporal dynamics in airport catchment areas: The case of the New York Multi Airport Region, Journal of Transport Geography, 90. https://doi.org/10.1016/j.jtrangeo.2020.102916.

Caset, F. & Marques Teixeira, F. (2020)
Visualizing the potential for transit-oriented development: An open and interactive planning support tool in Flanders, Belgium (under review)

Marques Teixeira, F., & Derudder, B. (2020)
Revealing route bias in air transport data: The case of the Bureau of Transport Statistics (BTS), Origin-Destination Survey (DB1B).
Journal of Air Transport Management, 82. https://doi.org/10.1016/j.jairtraman.2019.101745

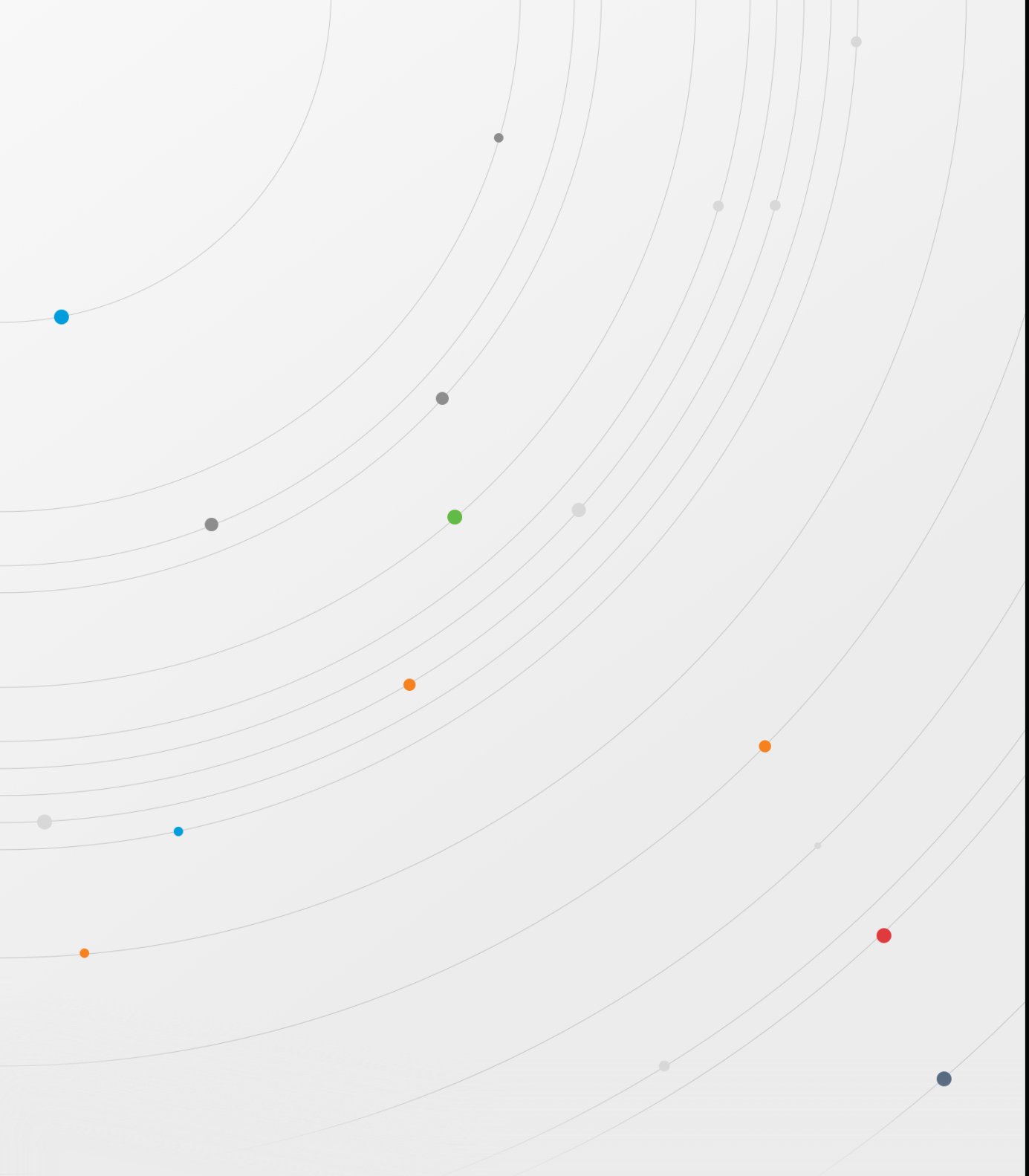
Caset, F., Marques Teixeira, F., Boussauw, K., Derudder, B., & Witlox, F. (2019)
Planning for nodes, places, and people in Flanders and Brussels: An empirical railway station assessment tool for strategic decision-making.
Journal of Transport and Land Use, Vol. 12, No. 1 (2019), pp. 811-837, http://dx.doi.org/10.5198/jtlu.2019.1483

Caset, F., Marques Teixeira, F., Boussauw, K., Derudder, B., & Witlox, F. (2019)
What strategies for which railway stations? An experiential approach to the development of a node-place based planning support tool in Flanders. In Proceedings of the BIVEC-GIBET Transport Research Days 2019. Ghent, Belgium.

Teixeira, F., & Derudder, B. (2018)
SKYNET: An R package for generating air passenger networks for urban studies. Urban Studies, 56(14). https://doi.org/10.1177/0042098018803258

Hardy Richter, Filipe M. Teixeira, Samira G. Ferreira, Ágnes Kittel, Attila Köfalvi, Beáta Sperlágh (2012)
Presynaptic α2-adrenoceptors control the inhibitory action of presynaptic CB1 cannabinoid receptors on prefrontocortical norepinephrine release in the rat,
Neuropharmacology,
Volume 63, Issue 5, Pages 784-797
https://doi.org/10.1016/j.neuropharm.2012.06.003.

Samira G. Ferreira, Filipe M. Teixeira, Pedro Garção, Paula Agostinho, Catherine Ledent, Luísa Cortes, Ken Mackie, Attila Köfalvi (2012)
Presynaptic CB1 cannabinoid receptors control frontocortical serotonin and glutamate release – Species differences,
Neurochemistry International,
Volume 61, Issue 2,
Pages 219-226



This dissertation deals with the need for, and the key role and development of data-driven tools and methodologies for minimizing data complexity in transport geography research. It does so by developing a diverse range of analytical tools that collectively show that this is both feasible and useful. In addition to developing these tools in the strict sense, the dissertation also examines various challenges associated with the process, ranging from acquiring data to data curation, analysis and visualization. The thesis is divided into four chapters: (1) I first introduce SKYNET, a flexible R package that allows generating bespoke air transport statistics for urban studies based on publicly available data from the BTS in the United States. (2) I explore how potential biases in air transport datasets can be revealed and detailed by focusing on the US Origin- Destination Survey (DB1B) and the Air Carrier Statistics-form 41 traffic (T-100) datasets. (3) I explore the spatio-temporal dynamics of airport catchment areas within the New York Multi Airport Region. (4) And finally, I present StationsRadar, a data-driven web-based tool developed to support integrated land use and transport strategy-making at railway station locations in the region of Flanders and the Brussels Capital Region.

FACULTY OF SCIENCES

GHENT UNIVERSITY

Following the tradition of elliptical orbit maps, each 'orbit' represents a month, and each 'object' represents an event in the four years of my research. When an orbit is missing, it simply represents a month without any particular events.